ELECTORAL FORENSICS:

TESTING THE "FREE AND FAIR" CLAIM

By

OLE J. FORSBERG

Bachelor of Science in Mathematics and Physics
University of Portland
Portland, OR
1990

Master of Science in Engineering in Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD
2010

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2014

ELECTORAL FORENSICS:

TESTING THE "FREE AND FAIR" CLAIM

Dissertation Approved:

Mark E. Payton

Dissertation Adviser

Joshua Habiger

Ye Liang

Jonathan Comer

Name: OLE J. FORSBERG

Date of Degree: MAY 2014

Title of Study: ELECTORAL FORENSICS: TESTING THE "FREE AND FAIR" CLAIM

Major Field: STATISTICS

Abstract: Democratic elections are the international norm for government legitimacy. Currently, election observers are the primary examiners of claims of democracy. Unfortunately, cost and access restrict their effectiveness. Electoral forensics complements these observers by statistically testing official results for evidence of violations of the democratic hypothesis.

This research seeks to revise statistical methods to increase their applicability and improve their statistical properties vis-à-vis the data types generated in elections. It focuses on three types of data typically reported by official agencies: candidate vote count, invalidation count, and geographical location.

Currently, the Benford test serves as the customary vote-count test. I improve this test by modifying it to consider the electoral division size. Similarly, present methods treat divisions as from a single population. I correct this by including a threshold (change-point) in the model. As such, the regression tests become more powerful, especially when used in conjunction with feasible generalized least squares regression. Finally, analyses tend to ignore the geographic nature of elections. I create the spatial-lag expansion model (SLEM) to compete with the popular geographically weighted regression (GWR) model.

My generalized Benford test improves the Benford test. However, the power is slight for two of the versions. Allowing for two populations in the election results increases the power of the regression tests while not affecting their sizes. Finally, SLEM betters GWR in terms of speed and power; however, the distribution of its test statistic must still be estimated using simulation.

Even with this advancement in the discipline, there remain a couple areas needing further treatment. First, GWR is attractive in that it allows modeling the data more closely. Further work should be done in devising an improved hypothesis-testing paradigm for it. Finally, electoral systems have unique aspects. Future research will be done in modeling electoral rules to create tests optimized for these unique conditions.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

June 12, 2009, saw Iranians vote for their next president. That night, after the ballots were counted, incumbent president Mahmoud Ahmadinejad was declared the winner with over 60% of the vote. The next day, protesters took to the streets to voice their belief that the election was fraudulent. Western governments concurred.

On November 28, 2010, Ivoirians cast ballots for president of this war-torn country, between incumbent president Laurent Gbagbo and opposition candidate Alassane Ouattara. Neither had received a majority of the vote in October's first round election, so they faced each other in the November runoff election. The Independent Electoral Commission declared Ouattara the winner of the second round. The Constitutional Court declared Gbagbo the winner. The two candidates rallied supporters, who took to the streets. Ouattara asserted the vote was free and fair and that the Independent Electoral Commission was the final arbiter. Gbagbo declared the vote free and fair and that the Constitutional Court was the final arbiter. Both candidates declared their own counts correct and their opponent's fraudulent. Thousands of civilians died in the resulting civil war.

On the same day, Haitians went to the polls to elect a parliament and a president. No candidate received a majority of the votes cast, so a runoff election was held on March 20, 2011. The two candidates receiving the greatest number of first-round votes were Mirlande Manigat (with 31%) and Jude Célestin (with 23%); they should have been the two candidates in the runoff. However, supporters of candidate Michel Martelly protested

the vote, forced a recount, denied the validity of the recount, and held demonstrations. Martelly and his supporters claimed election fraud.

1.1. Raison d'Être

The events described in these three vignettes are not unique. Elections happen. Frequently, losing candidates claim the vote fraudulent. Many elections have international groups observing, watching for electoral irregularities. These groups are not in all voting precincts or in all counting locations, thus fraud may still go unseen. Furthermore, these groups are unable to see systematic unfairnesses in the electoral system. Even with international observers, fraud exists. Unfair electoral systems exist. Violations of the democratic claim exist.

This is the *raison d'être* behind pursuing this research. At their most abstract, elections are observed outcomes of the sums of categorical random variables. Random variables have distributions, expected values, expected correlations, etc. This research scours several areas of statistics, collecting and modifying multiple statistical techniques in order to better detect violations of democratic claims—violations of the free and fair hypothesis (Guterres 2008).

1.2. Testing the Free and Fair Hypothesis

There is no single definition of a free and fair election. Political Scientists spend much ink on the precise meanings of these two terms. At the intersection of all definitions may be that a free and fair election is one in which each citizen's vote counts the same (Kirkpatrick 1984; Wantchekon 1999).

This definition has testable implications. First, ballot box stuffing is a violation. Those who stuff the box cast more ballots and have ballots more apt to be correctly filled out and counted than do those of other citizens. Second, declaring a ballot invalid must

be done independent of the voter's age, gender, education level, ethnicity, etc., or for whom the ballot is cast.

Testing the first implication is easily done with the vote counts and the distribution of vote counts in a free and fair election. The former is frequently reported by the countries. The latter is discussed in Chapter 2. While each ballot cast *can* be modeled as a Bernoulli random variable, the ballot count totals are neither Binomial nor Hypergeometric random variables. However, Simon Newcomb (1881) and Frank Benford (1938) posited a distribution for unbounded counts, which Walter Mebane (2010) used to examine the 2009 Iranian election.

In Chapter 2, I closely examine the Benford test, including its assumptions. In addition to the Benford test, I examine several related statistical techniques that rely solely on the distribution of reported vote counts at the electoral division level. The results are interesting, because this type of test offers the best opportunity to test for election fraud.

Testing the second implication is also easily done—as long as the data are available. Such tests are variations on regressing the invalidation rate against one or more demographic variables or the level of candidate support (or both). The former regressions may be able to detect unintended unfairness in the election. The latter regressions may be able to detect electoral fraud. In Chapter 3, I examine several regression methods.

These models range from the venerable ordinary least squares (OLS) regression to the much newer feasible generalized least squares (FGLS) regression and Bayesian regression. Each method has strengths: OLS is easily performed. FGLS allows one to estimate the arbitrary correlation matrix. Bayesian regression produces posterior distributions of the population parameters, not just a final estimate. Each regression model also has weaknesses.

I compare the several regression methods in terms of the observed Type I Error rate, and the categorization methods in terms of mean square error.

All of the regression tests explored in Chapter 3 make some assumption regarding the independence of measurements and the constancy of effects. As voting is an *inherently* spatial event, it may be that the errors are spatially correlated or that the parameters are spatially varying. The regression techniques of Chapter 3 cannot cover such cases as stated.

In Chapter 4, I explore the problem of space and propose several solutions. While controlling the spatial correlation is an option, spatially varying effects are interesting in and of themselves. As such, merely controlling the effect of space is fundamentally unsatisfying. Currently, there are three methods for modeling in the presence of spatial correlation: spatially-lagged dependent variable regression, Casetti's expansion method, and Fotheringham's geographically weighted regression. All have strengths and weaknesses.

The spatially lagged dependent variable model is easy to implement. It merely includes a spatially-lagged dependent variable as an independent variable. That this method can be used in all of the regression methods of Chapter 3 is its strength. That its effect estimates are biased is its weakness. However, is the bias small enough as to allow its flexibility to outweigh it? Furthermore, while the estimates are biased, is its mean square error smaller than that of other methods? If so, then this method may still be preferred.

Emilio Casetti (1972) created the expansion method to model spatially varying effects. It is a type of regression and can also work within the context of the Chapter 3 regression methods. This method includes functions (linear, quadratic, logistic, etc.) of the location as additional independent variables and of currently used independent variables. Its main weakness is that the function must be known *a priori*.

A. Stewart Fotheringham and associates (1996; 1997; 1998) developed a flexible alternative to Casetti's expansion method. In lieu of the researcher specifying the shape of the effect surface, Fotheringham suggested allowing the data to determine the shape.

As such, he and his associates created the geographically weighted regression (GWR) method. In this method, the effects are estimated at each point in space using the nearby points as data, usually weighted according to distance. The major drawback is that there is no parametric test either for the effects or for whether those effects are spatially varying. This severely limits its utility.

In Chapter 4, I then compare the GWR method with my proposed method (SLEM) in terms of the Type I Error rate and of the power. I conclude by discussing a few interesting points about the power curves.

While each chapter includes limited illustrations of the techniques discussed, I fully apply all of the approved techniques discussed to two cases. In Chapter 5, I apply them to a questionable election, the Unity referendum of 2011 in Southern Sudan. In Chapter 6, I apply them to an election that is allegedly free and fair, the 2008 US Presidential election in Colorado. These two chapters should illustrate a correct application of the methods.

1.3. Simulating Elections

Throughout this research, there is a need to compare estimators. There are several available criteria. The bias is a measure of how close the average estimate is to the true value. The mean square error is a measure of how concentrated the estimates are around the true value. Both of these are important in comparing estimators. However, before an estimator should be compared, it should meet one criterion: The true Type I Error rate must be close to the nominal rate. Testing this is rather unambiguous. One generates multiple "free and fair" elections, performs the test on each, the compares the rejection rate with the nominal rate, α. This is easy if you can generate free and fair elections. Deckert et al. (2011) assert that there is no *a priori* distribution for voting.

The second aspect of a test that should be examined is its power—the ability of the test to reject a false null hypothesis. Where the previous requires randomly generated free

and fair elections, this requires randomly generated *unfair* elections. This is significantly more difficult. If there is no distribution for a fair election, then it cannot be modified to be unfair.

To solve these issues, I do two things. First, I make the explicit assumption that the 2008 presidential elections in the United States are free and fair. From those election data, I extract the division size and the proportion of the vote in each division cast for John McCain—the candidate of the incumbent party. The former variable is the electoral division sizes; the latter, the proportion data.

The free and fair elections are generated by drawing a random sample, with replacement, from the proportion data. Those proportions then multiply the electoral division size data, which is not permuted. As neither the electoral division sizes nor the proportion of the vote in favor of the candidate are fraudulent, this method has face validity.

Generating unfair elections is not as easy; elections can be unfair in several ways. In Chapter 2, I modify the fair elections with different levels of three contaminants. Contaminant One is shifting the leading digit up one. As this chapter consists of digit tests, this is equivalent to ballot counters writing in fraudulent counts with the next leading digit. The second contaminant is the uniform distribution. This would represent the same situation as above, but with the ballot counter replacing the true counts with counts having a random leading digit. The third contaminant is replacing all initial digits with a '9.'

For Chapter 3, the leading digits are not important, *per se*. What is important is the relationship between the invalidation rate and the candidate support in each division. Under the null hypothesis, this relationship is independent, which can be easily simulated. Again, the contaminated elections are more difficult to simulate.

These fraudulent elections are generated by selecting a threshold ($\tau = 0.60$), creating a sloped regression line to its right, then add varying levels of "noise" (increase

the variance of the residuals). When the noise is great, the mean square error of the tests should be high. When the noise is minor, the MSE should be small.

This method carries over without change into Chapter 4, where geography becomes an important factor. I do not explicitly consider geography when placing the stuffing ballots; I allow the underlying geographic correlation of candidate support do that.

1.4. Notation

Finally, before beginning our journey, we need to discuss notation briefly. Rarely is there standard notation in statistics. As such, I shall provide a listing of notation I use throughout this research.

The following are several distributions used in this research.

Name	**Symbol**	**Parameter(s)**	**Page**
Benford	$\mathfrak{B}_1$	θ	15
Beta	BETA	α, β	84
Gamma	GAMMA	κ, θ	85
Log-uniform	$\mathcal{LU}$	θ_a, θ_b	14
Logit-normal	$Lgt\mathcal{N}$	μ, σ^2	37
Multinomial	$\mathcal{M}ulti$	n, $\boldsymbol{\pi}$	30
Normal	$\mathcal{N}$	μ, σ^2	37
Uniform	$\mathcal{U}$	θ_a, θ_b	14

The following indicate the meaning of the symbol based on its size and national origin.

Majuscule Greek	Parameter space	Θ
Minuscule Greek	Population parameter	θ
Majuscule Roman	Random variable	X
Minuscule Roman	Observed random variable or non-random variable	x

Finally, $\mathbf{X}$ is a matrix or a vector, X is a scalar, "ln" is the natural logarithm (base-e), and "log" is the common logarithm (base-10).

1.5. Conclusion

This introductory chapter framed the issue: Some elections experience fraud and the current detection techniques are few and far between. Statistical tests require data. Increased amounts of data provide a better chance of detecting electoral fraud and other violations of the free and fair hypothesis.

Unfortunately, those data are controlled by the same government we hope to test. As such, the data may be sparse. For instance, the 2009 Afghan presidential election only reported the vote counts in each division (*wilāyat*). These data require a different battery of tests than elections in which vote counts, invalidation counts, and demographic variables are measured to the precinct level.

In the next chapter, I introduce you to the 2009 Afghan election as well as digit tests that can test such data.

CHAPTER 2

DIGIT TESTS

Afghanistan, 2009. Afghanistan's second presidential election under its 2004 constitution took place on August 20, 2009. This election pitted incumbent Hamid Karzai against challenger Abdullah Abdullah, M.D., of the United National Front. The campaign saw the use of the state-run media and intimidation to affect the outcome. Public opinion polls gave Karzai a wide lead, but falling short of giving him the needed majority to avoid an October runoff election.

Election day was predictable. The Taliban called for a boycott of the "Western-led" poll. Violence broke out at several polling places. Challengers made charges of ballot-box stuffing and of false counting (Galbraith 2009).

Two days later, the Independent Election Commission (IEC) announced its official results: Karzai received 54.6% of the valid votes cast (Afghanistan 2009b). The UN-dominated Electoral Complaints Commission (ECC), the final arbiter of the election, refused to validate the election until all fraud claims were adjudicated. However, even this did not progress smoothly. Methods of determining the proportion of votes to remove met with charges of Western meddling.

One thing was true: If there were a second round election, it would be between Karzai and Abdullah. Under US pressure, Karzai accepted a runoff election, now to be scheduled for November 7, 2009. Abdullah demanded certain members of the IEC be removed. They refused, Abdullah withdrew from the runoff election, and the IEC named Hamid Karzai President-elect (Galbraith 2009).

Doubts over the extent of the fraud claims linger. However, because of the extant

situation, the only numbers easily available are the vote counts at the province level (*wilāyat*).

2.1. Introduction

Across the world, democratic elections are becoming the norm. Several non-democracies even hold elections to provide a veneer of popular legitimacy (Wantchekon 1999). As such, the presence of the word 'democracy' does not necessarily indicate a democratic election. While Political Scientists do not agree on the specific requirements of a democratic election, they do appear to agree that electoral fraud is not democratic. When the followers of a candidate stuff a ballot box with prepared ballots or falsify vote counts, fraud has taken place. Luckily, both of these acts leave behind evidence. The trick is to find that evidence.

A goal of electoral forensics is to determine, using statistical methods, if an election violates the assumption of democracy, if the election violates the free and fair hypothesis. Democratic elections must be free and fair. They must be free in the sense that voting is allowed. They must be fair in the sense that a person's ballot counts the same an any other person's ballot—or at least has the same *probability* of counting.

Unfortunately, the information needed to test for fraud is controlled by the very government being tested. This usually means the level of information available is quite limited. With this reality, electoral forensics seeks testing methods given the slight information available. Fortunately, many elections provide the vote counts for the candidates in each first-level administrative division. This is what Afghanistan's Independent Election Commission offered from the 2009 Presidential election.

Currently, the Benford test is the standard method for testing for violations of the free and fair hypothesis when only vote counts are available (Mebane 2010). The Benford test compares the actual digit frequencies to a hypothesized distribution. A usual method

of frequentist statistics is to calculate a test statistic for which a distribution is knowable. From that, a p-value can be calculated. This is the strength of the Benford test. It offers a statistical test of the free and fair hypothesis based solely on the assumed distribution of vote counts.

It is not the only possible test, nor is it the best in all situations. This chapter introduces the history of the Benford test and its uses. It then explores its applicability to election data, concluding that the extant test has severe shortcomings.

To compensate for those issues, I modify it and formulate additional tests. Those tests include both parametric and non-parametric simulation. To accomplish a non-parametric simulation test, I introduce the Logit-normal distribution, which models the vote proportion for each candidate in an election. With this distribution, the parametric test surpasses both the Benford test and the non-parametric simulation test.

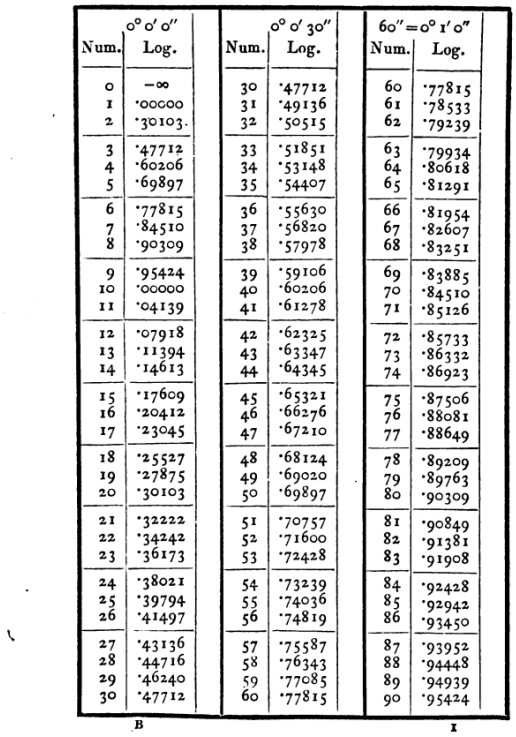
2.2. History

While browsing a book of logarithms (e.g., Figure 2.1), Harvard astronomer Simon Newcomb (1881, p. 39) noticed something interesting:

> That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones. The first significant figure is oftener 1 than any other digit, and the frequency diminishes up to 9.

From this observation, he derived the distribution of those leading digits. More importantly to him, he concluded that one could distinguish between a table of numbers arising in nature and their logarithms. The leading digits of the former follow the described distribution; of the latter, a uniform distribution.

Fifty-two years later, Physicist Frank Benford made the same observation, "[t]he pages containing the logarithms of the low numbers 1 and 2 are apt to be more stained



Num.	0° 0′ 0″ Log.	Num.	0° 0′ 30″ Log.	Num.	60″ = 0° 1′ 0″ Log.
0	−∞	30	·47712	60	·77815
1	·00000	31	·49136	61	·78533
2	·30103	32	·50515	62	·79239
3	·47712	33	·51851	63	·79934
4	·60206	34	·53148	64	·80618
5	·69897	35	·54407	65	·81291
6	·77815	36	·55630	66	·81954
7	·84510	37	·56820	67	·82607
8	·90309	38	·57978	68	·83251
9	·95424	39	·59106	69	·83885
10	·00000	40	·60206	70	·84510
11	·04139	41	·61278	71	·85126
12	·07918	42	·62325	72	·85733
13	·11394	43	·63347	73	·86332
14	·14613	44	·64345	74	·86923
15	·17609	45	·65321	75	·87506
16	·20412	46	·66276	76	·88081
17	·23045	47	·67210	77	·88649
18	·25527	48	·68124	78	·89209
19	·27875	49	·69020	79	·89763
20	·30103	50	·69897	80	·90309
21	·32222	51	·70757	81	·90849
22	·34242	52	·71600	82	·91381
23	·36173	53	·72428	83	·91908
24	·38021	54	·73239	84	·92428
25	·39794	55	·74036	85	·92942
26	·41497	56	·74819	86	·93450
27	·43136	57	·75587	87	·93952
28	·44716	58	·76343	88	·94448
29	·46240	59	·77085	89	·94939
30	·47712	60	·77815	90	·95424

B 1

Figure 2.1: *A scan of a page from Richard Farley's* Tables of Logarithms *(1839). This is the first table in the book. Its level of wear can be inferred through the poor quality of the top edge of the book.*

and frayed by use than those of the higher numbers 8 and 9" (Benford 1938, p. 551). Benford's contribution is not the observation, but the application (Benford 1938, p. 551):

> . . . no one could be expected to be greatly interested in the condition of a table of logarithms, but the matter may be considered more worthy of study when we recall that the table is used in the building up of our scientific, engineering, and general factual literature. There may be, in the relative cleanliness of the pages of a logarithm table, data on how we think and how we react when dealing with things that can be described by means of numbers.

In the course of his research, he gathered numbers from many sources and tested the leading-digit distribution for each source and for the union of the sources. These sources ranged from lengths of rivers to addresses in the telephone book to the mathematical sequence $\{n, n^2, n^3, \ldots\}$. While the individual sources did not always follow the prescribed distribution (river lengths did, the mathematical sequence did not), the entirety of the sources did.

Examining the table, Benford concluded that those sources of a random nature followed the "logarithmic law" much more closely than those sources arising from a mathematical formula (Benford 1938, p. 557):

> These facts lead to the conclusion that the logarithmic law applies particularly to those outlaw numbers that are without known relationship rather than to those that individually follow an orderly course; and therefore the logarithmic relation is essentially a Law of Anomalous Numbers.

With that said, in the third part of the article, Benford derives the distribution of leading digits when the numbers are integers from 1 to a given upper bound. That such a sequence "follows an orderly course" should indicate that the leading digits would not well follow the Law of Anomalous Numbers. However, Benford shows this sequence closely agrees with this "Law" for upper bounds larger than 1000.

It is this distribution of leading digits of the integers from 1 to n that became the basis for using statistics to detect fraud (Cho and Gaines 2007; Hill 1995; Mebane 2010; Nigrini 2011, 2012) and deviations from random behavior (Carslaw 1988; Ley 1996).

2.2.1 Derivation of the Benford distribution. To calculate Benford's probability mass function, we return to Newcomb. Recall that Newcomb observed the earlier pages of logarithm books were more worn than later pages; that is, those pages dealing with numbers beginning with a 1 were used more often. From this, Newcomb

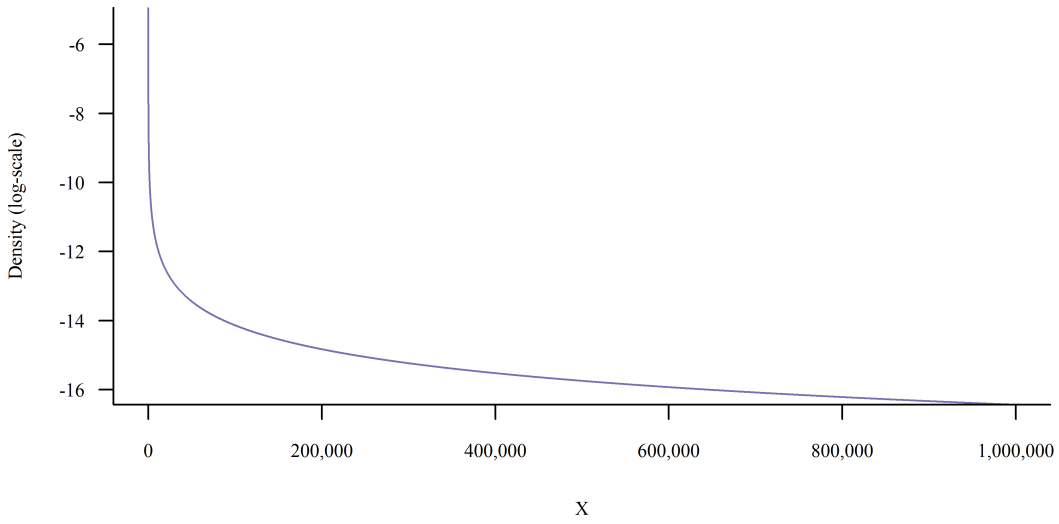


Figure 2.2: *The probability density function of the* $\mathcal{LU}(0,6)$ *distribution over its support,* $[1, 10^6]$

deduced that the probability of a given digit being the leading digit of a number was $\log(d+1) - \log(d)$.

Definition 2.1 (Log-uniform Distribution)**.** *Define* Y *as a Uniformly distributed random variable,* $Y \sim \mathcal{U}(0,\theta)$*, with* θ *being its maximum value. Then* $X := 10^Y$ *has a Log-uniform distribution, symbolized as* $X \sim \mathcal{LU}(0,\theta)$*, with support* $X \in [1, 10^\theta]$*.*

This distribution has a cumulative distribution function $F_X(x) = \frac{\log x}{\theta}$ and probability density function $f_X(x) = \frac{1}{x\ \theta\ \ln 10}$, where log and ln are the common and natural logarithm functions, respectively. Figure 2.2 provides the probability density function for the Log-uniform distribution over its support; Figure 2.3, its cumulative distribution function.

Definition 2.2 (Leading-digit Function)**.** *Define the leading-digit function* $\mathcal{D}_1(x)$ *taking a non-negative real number as its argument and returning the first (leading) digit of that number:*

$$\mathcal{D}_1(x) := \begin{cases} \left\lfloor x\ 10^{-\lfloor \log x \rfloor} \right\rfloor & x > 0 \\ 0 & x = 0 \end{cases}$$

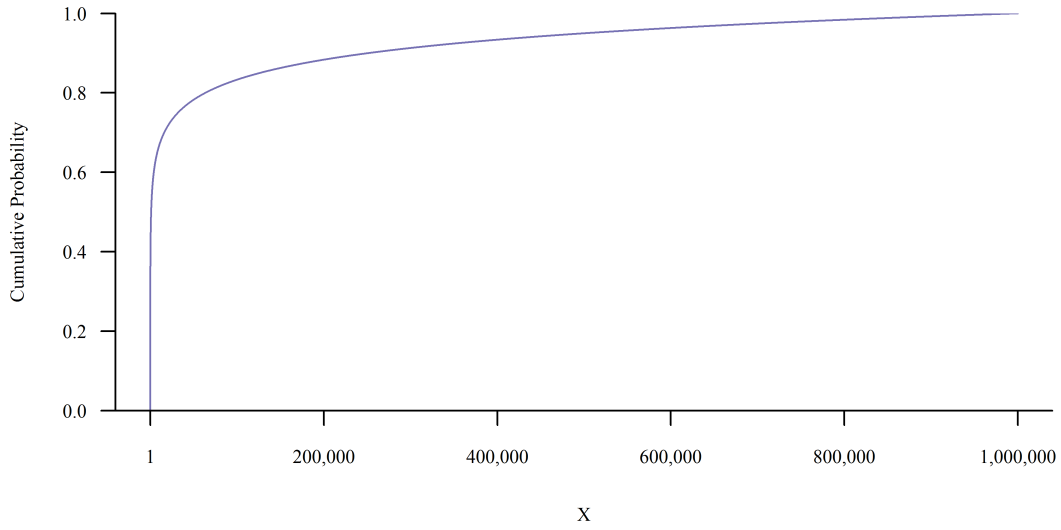


Figure 2.3: *The cumulative distribution function of the* $\mathcal{LU}(0, 6)$ *distribution over its support set,* $\mathcal{S} = [1, 10^6]$.

This leads to the Benford distribution:

Definition 2.3 (Benford Distribution)**.** *If* $X \sim \mathcal{LU}(0, \theta)$*, then the distribution of* $\mathcal{D}_1(X)$ *is the Benford distribution of order 1, designated* $D \sim \mathfrak{B}_1(0, \theta)$*, where* $D := \mathcal{D}_1(X)$.

If $\theta \in \mathbb{N}$, then $\mathfrak{B}_1(0, \theta)$ is termed the integer Benford distribution. Now, to determine the probability mass function of the Benford distribution, we have two cases. In the first, $\theta \in \mathbb{N}$. In the second, $\theta \notin \mathbb{N}$. The derivation in the first case is simple; in the second, more complex.

Possibility 1: $\theta \in \mathbb{N}$: Let us now deal with the first possibility. Here, $\theta \in \mathbb{N} := \{1, 2, 3, 4, \ldots\}$. This is the simple case as the probability function of the Benford distribution is independent of θ.

Newcomb (1881) provided a proof of this based on a limiting equispaced circular distribution. Benford's proof (1938) relies on a counting argument. The following proof relies on the cumulative distribution function.

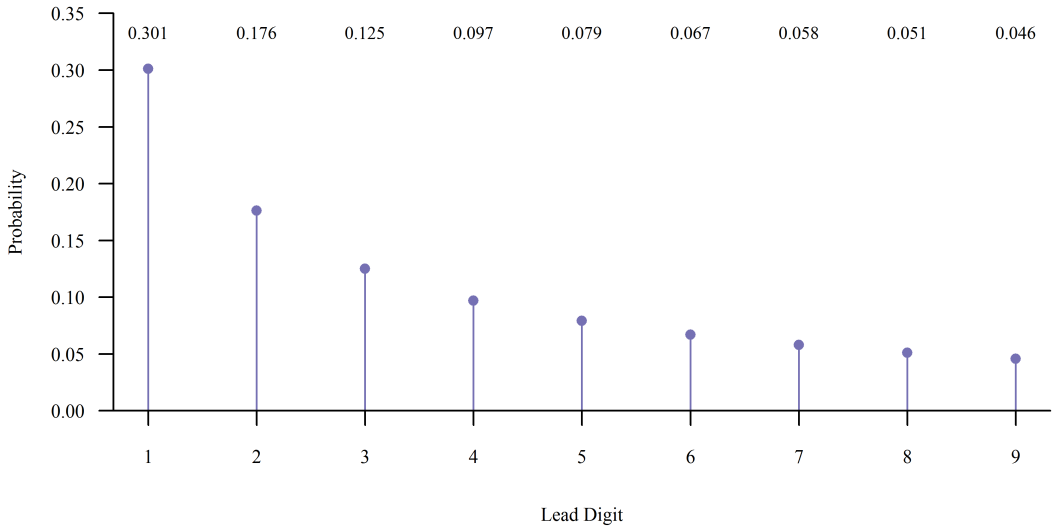


Figure 2.4: *Distribution of lead digits according to the integer Benford distribution,* $\mathfrak{B}_1(0,\theta)$*,* $\theta \in \mathbb{N}$*. Note that the lead digit is not uniformly distributed. In fact, the probability of the lead digit being a '1' is more than six times that of it being a '9'.*

Lemma 2.4. *The probability mass function of the integer Benford distribution,* $\mathfrak{B}_1(0,\theta) := \mathbb{P}[\mathcal{D}_1(X) = d\,|\,\theta]$*, is independent of* θ *and is* $\log\left(\frac{d+1}{d}\right)$ *for* $\theta \in \mathbb{N}$*.*

Proof. Recall $X \sim \mathcal{LU}(0,\theta)$ and $F_X(x) = \frac{\log x}{\theta}$. Then, $\mathbb{P}[\mathcal{D}_1(X) = d\,|\,\theta]$ is equivalent to calculating the area under the pdf curve corresponding to values with leading digit d. Equivalently, this is summing up over integer powers of 10 the differences in the CDF values of $(d+1)\times 10^i$ and $d \times 10^i$. That is:

$$\begin{aligned}
\mathbb{P}[\mathcal{D}_1(X) = d\,|\,\theta] &= \sum_{i=0}^{\theta-1} \Big(F_X\big((d+1)\times 10^i\big) - F_X\big(d \times 10^i\big) \Big) \\
&= \sum_{i=0}^{\theta-1} \left(\frac{\log(d+1)+i}{\theta} - \frac{\log d + i}{\theta} \right) \\
&= \sum_{i=0}^{\theta-1} \left(\frac{\log(d+1)}{\theta} - \frac{\log d}{\theta} \right) \\
&= \log(d+1) - \log d
\end{aligned}$$

Algebra provides the conclusion. □

Statistic	Definition	Approx Value
$\mathbb{E}[D]$	μ_1	3.440
$\mathbb{V}[D]$	$\mu_2 - \mu_1^2$	6.057
Skew(D)	μ_3/σ^3	0.796
Kurtosis(D)	μ_4/σ^4	22.444

Table 2.1: *Approximate values of some population statistics of the* $\mathfrak{B}_1(0,\theta)$ *distribution, with* $\theta \in \mathbb{N}$

The probability mass function (pmf) of the integer Benford distribution, $\mathfrak{B}_1(0,\theta)$, is graphed in Figure 2.4. Note the differences in probabilities among the digits.

Proposition 2.5. *Let* $D \sim \mathfrak{B}_1(0,\theta)$*, with* $\theta \in \mathbb{N}$*. Table 2.1 holds.*

These simple population statistics are of little help by themselves. A symmetric 95% confidence interval on a single leading digit is $[1, 9]$. However, multiple samples shrink the width of the confidence interval, making testing possible.

Proposition 2.6. *Let* $D_i \overset{iid}{\sim} \mathfrak{B}_1(0,\theta)$*, with* $i \in \{1, 2, \ldots, n\}$*. If* $X := \sum_{i=1}^{n} D_i$*, then* $\mathbb{E}[X] = \mathbb{E}[D_1]$ *and* $\mathbb{V}[X] = \frac{1}{n}\mathbb{V}[D_1]$*.*

Proposition 2.6 is useful in estimating confidence intervals for the average leading digit in a dataset. For instance, the Afghan dataset contains vote counts for each candidate from the 34 *wilāyat*. If the leading digits for each *wilāyat* follow the Benford distribution, then the expected leading digit will be 3.44 and an estimated 95% confidence interval for $n = 34$ will be from 2.65 to 4.29, using simulation. Using a Normal approximation (via the Central Limit Theorem), the symmetric 95% confidence interval will be from 2.61 to 4.27. The observed mean leading digit for President Hamid Karzai is 3.294118, which is within both confidence intervals. Thus, on the basis of this test, there is no significant evidence of electoral fraud in the 2009 Afghan presidential election.

POSSIBILITY 2: $\theta \notin \mathbb{N}$: The first case was the familiar result from Newcomb (1881), Benford (1938), and Mebane (2010). Newcomb (1881) and Benford (1938) allude to this second (more general) case, but neither explored nor derived it.

Theorem 2.7. *The probability mass function of* $\mathfrak{B}_1(0,\theta) := \mathbb{P}[\mathcal{D}_1(X) = d\,|\,\theta]$ *is*

$$\frac{\lfloor\theta\rfloor}{\theta}\cdot\log\left(\frac{d+1}{d}\right)+\begin{cases}0 & 10^{\theta-\lfloor\theta\rfloor} < d\\ \frac{1}{\theta}\Big(\theta-\lfloor\theta\rfloor+\log\left(\frac{1}{d}\right)\Big) & d \le 10^{\theta-\lfloor\theta\rfloor} < d+1\\ \frac{1}{\theta}\log\left(\frac{d+1}{d}\right) & d+1 \le 10^{\theta-\lfloor\theta\rfloor}\end{cases}$$

and is continuous in θ.

Proof. With respect to the function, there are three cases. In the first case, the leading digit of the upper bound of X is less than the digit under consideration. In the second case, the leading digit of the upper bound of X *is* the digit under consideration. In the third case, the leading digit of the upper bound of X is greater than the digit under consideration.

Case 1: In this case, the leading digit of the upper bound of X is less than the digit under consideration.

$$\begin{aligned}\mathbb{P}[\mathcal{D}_1(X) = d\,|\,\theta] &= \sum_{i=0}^{\lfloor\theta\rfloor-1} F\Big((d+1)\times 10^i\Big) - F\Big(d\times 10^i\Big)\\ &= \lfloor\theta\rfloor\left(\frac{1}{\theta}\Big(\lfloor\theta\rfloor+\log(d+1)\Big)-\frac{1}{\theta}\Big(\lfloor\theta\rfloor+\log d\Big)\right)\\ &= \frac{\lfloor\theta\rfloor}{\theta}\left(\Big(\lfloor\theta\rfloor+\log(d+1)\Big)-\Big(\lfloor\theta\rfloor+\log d\Big)\right)\\ &= \frac{\lfloor\theta\rfloor}{\theta}\log\left(\frac{d+1}{d}\right)\end{aligned}$$

Case 2: In this case, the leading digit of the upper bound of X *is* the digit under consideration.

$$\begin{aligned}\mathbb{P}[\mathcal{D}_1(X) = d\,|\,\theta] &= \sum_{i=0}^{\lfloor\theta\rfloor-1} F\Big((d+1)\times 10^i\Big) - F\Big(d\times 10^i\Big) \\ &\qquad + F\Big(10^\theta\Big) - F\Big(d\times 10^{\lfloor\theta\rfloor}\Big) \\ &= \frac{\lfloor\theta\rfloor}{\theta}\log\left(\frac{d+1}{d}\right) + F\Big(10^\theta\Big) - F\Big(d\times 10^{\lfloor\theta\rfloor}\Big) \\ &= \frac{\lfloor\theta\rfloor}{\theta}\log\left(\frac{d+1}{d}\right) + \frac{\theta}{\theta} - \frac{\lfloor\theta\rfloor + \log d}{\theta} \\ &= \frac{\lfloor\theta\rfloor}{\theta}\log\left(\frac{d+1}{d}\right) + 1 - \frac{\lfloor\theta\rfloor}{\theta} - \frac{\log d}{\theta}\end{aligned}$$

Case 3: In this case, the leading digit of the upper bound of X is greater than the digit under consideration.

$$\begin{aligned}\mathbb{P}[\mathcal{D}_1(X) = d\,|\,\theta] &= \sum_{i=0}^{\lfloor\theta\rfloor-1} F\Big((d+1)\times 10^i\Big) - F\Big(d\times 10^i\Big) \\ &\qquad + F\Big((d+1)\times 10^{\lfloor\theta\rfloor}\Big) - F\Big(d\times 10^{\lfloor\theta\rfloor}\Big) \\ &= \frac{\lfloor\theta\rfloor}{\theta}\log\left(\frac{d+1}{d}\right) + F\Big((d+1)\times 10^{\lfloor\theta\rfloor}\Big) - F\Big(d\times 10^{\lfloor\theta\rfloor}\Big) \\ &= \frac{\lfloor\theta\rfloor}{\theta}\log\left(\frac{d+1}{d}\right) + \frac{\lfloor\theta\rfloor\log(d+1)}{\theta} - \frac{\lfloor\theta\rfloor\log d}{\theta} \\ &= \frac{\lfloor\theta\rfloor}{\theta}\log\left(\frac{d+1}{d}\right) + \frac{1}{\theta}\log\left(\frac{d+1}{d}\right)\end{aligned}$$

The Theorem's formula follows. Continuity follows from the function values at the end points being equal. □

If $\theta \in \mathbb{N}$, Theorem 2.7 reduces to Lemma 2.4, as it should. To see this, note that $\theta = \lfloor\theta\rfloor$ when $\theta \in \mathbb{N}$ and that $10^0 \leq d$ for all $d \in \{1, 2, \ldots, 9\}$. The result follows from being in Case 1 in the Lemma.

Both Newcomb and Benford proved their results in the case where the values of X were unbounded; i.e. when $\theta \to \infty$. Using Theorem 2.7, we can now prove their assertions differently.

Corollary 2.8. *As $\theta \to \infty$, the generalized Benford distribution converges in distribution to the integer Benford distribution; that is, $\mathfrak{B}_1(0,\theta) \xrightarrow{\mathfrak{L}} \mathfrak{B}_1(0,n)$, $n \in \mathbb{N}$.*

Proof. From Theorem 2.7, $\mathbb{P}[\mathcal{D}_1(X) = d]$ varies between a lower bound at $d = 10^{\theta - \lfloor\theta\rfloor}$ and an upper bound at $d + 1 = 10^{\theta - \lfloor\theta\rfloor}$.

For the lower bound, substitution gives:

$$\begin{aligned}
\text{Lower bound} &= \frac{\lfloor\theta\rfloor}{\theta} \log\left(\frac{d+1}{d}\right) + \left(1 - \frac{\lfloor\theta\rfloor}{\theta} - \frac{\theta - \lfloor\theta\rfloor}{\theta}\right) \\
&= \frac{\lfloor\theta\rfloor}{\theta} \log\left(\frac{d+1}{d}\right) \\
&\to \log\left(\frac{d+1}{d}\right), \text{ as } \theta \to \infty.
\end{aligned}$$

Similarly, the upper bound is achieved at $d + 1 = 10^{\theta - \lfloor\theta\rfloor}$. Substitution gives

$$\begin{aligned}
\text{Upper bound} &= \frac{\lfloor\theta\rfloor}{\theta} \log\left(\frac{d+1}{d}\right) + \frac{1}{\theta} \log\left(\frac{d+1}{d}\right) \\
&= \frac{\lfloor\theta\rfloor + 1}{\theta} \log\left(\frac{d+1}{d}\right) \\
&\to \log\left(\frac{d+1}{d}\right), \text{ as } \theta \to \infty.
\end{aligned}$$

Since the upper bound converges to the lower bound, the Two Policemen and a Drunk Theorem (a.k.a. the Squeeze Theorem) tells us that $\mathfrak{B}_1(0, \theta) \xrightarrow{\mathfrak{L}} \mathfrak{B}_1(0, n)$, $n \in \mathbb{N}$, as $\theta \to \infty$. □

Figure 2.5 provides a graphical display of the leading digit probabilities for values of θ through $\theta = 6$. The top graph demonstrates how the mean leading digit varies with the value of θ. Note that $\mathbb{E}[D] \to 3.440$ as $\theta \to \infty$.

Now that we have a sufficient grounding for the Benford distribution, it is time to turn to its current limited use in electoral forensics.

2.3. The Benford test

Many fraud tests, such as the Benford test, are based on the assumption that humans are either not random, or are incorrectly random. For example, were someone to provide a listing of 1000 'H' and 'T' values, it would be straightforward to test if the data resulted

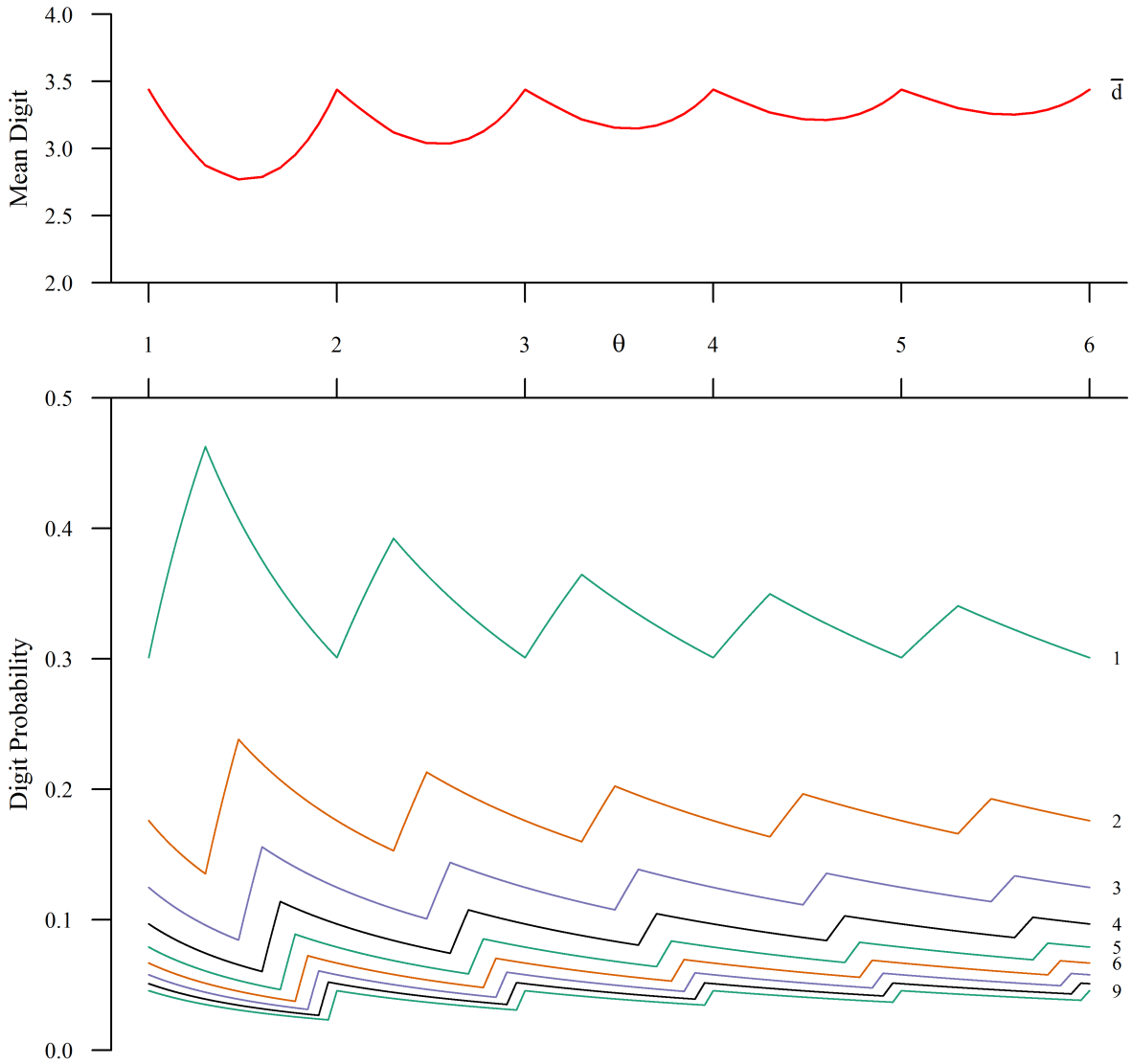


Figure 2.5: *Means and probabilities for various values of θ. The top plot is the mean leading digit; the bottom, leading digit probabilities with probabilities for the digit 1 at top, and the rest in decreasing order.*

from a thousand coin flips or if someone merely attempted to *physically* simulate 1000 coin flips. Aspects we could test on the coin flip data would include the numbers of 'H' entries, the average run length, and the distribution of run lengths. Under the null hypothesis that the data were generated from a fair coin, we know the distribution of each of the statistics.

As such, we can determine the probability of observing such data were the null hypothesis true. Were the data generated under the alternative hypothesis that they were generated "randomly" by a human, we know from experience that the distribution of the number of Heads is much more leptokurtotic (peaked) than it would be under the null hypothesis and that the average run length will be much shorter.

This is the idea behind digit tests in electoral forensics. Election results, while not coin flips, should also demonstrate some divergence from the expected null distribution if they are tainted by vote-count fraud—by humans recording knowingly false vote counts

How can vote-count fraud take place? While all electoral systems are unique in their specific structures, generalities can be made. Frequently, the votes are not counted in a single, central location. Votes cast at the precinct level tend to be counted at the precinct level, with vote counts forwarded to the national electoral commission, where the totals are counted and the final announcements are made. The ballot papers are later shipped to the national electoral commission (if at all).

Ghana follows this strategy. When the polling station closes, counting begins there, in full view of candidate agents. When the ballots are counted, the totals are recorded in the "Blue Book" for that polling station, with the candidate agents signing off on the legitimacy of the count and receiving copies of the counts. Those vote counts are then forwarded to the Electoral Commission of Ghana in Accra, which collects them, totals them, and announces the official winners (Ghana 2012, §35).

Conversely, Afghanistan does not follow this process. In 2009, the polling places sent all ballot papers to a central counting facility in Kabul (Afghanistan 2009a, Article 12). In such cases, it is much easier to perpetrate vote-count fraud; one merely has to control the small group doing the centralized counting.

In countries like Afghanistan, changing the counts (or even creating false counts) is relatively easy. The digit most likely changed is the lead digit, as that digit has the most impact on the election. As mentioned above, rarely are humans correctly random.

They will attempt to force the appearance of randomness to hide their tracks, but they will use the wrong distribution. That is, the distribution of vote counts in fraudulent precincts is different than the distribution of vote counts in non-fraudulent districts.

The current digit test is the (integer) Benford test (Hill 1995; Mebane 2010). According to the integer Benford distribution, the distribution of those first digits is

$$\mathbb{P}[\mathcal{D}_1(X) = d] = \log\left(\frac{d+1}{d}\right),$$

where $d \in \{1, 2, \ldots, 9\}$ is the leading digit (see Lemma 2.4).

This distribution is used in forensic accounting to detect fraud in expense accounts (Carslaw 1988; Cho and Gaines 2007; Nigrini 2011, 2012); in elections, to detect vote-counting fraud (Mebane 2010). Theoretically, the distribution of the leading digit of votes for a given candidate follows the integer Benford distribution. As humans tend to randomize much more uniformly, vote-count fraud will produce a distribution different from the Benford distribution. As such, a simple chi-squared test can be (and has been) used to detect this type of election fraud (Carslaw 1988).

2.3.1 The Violated Assumption. However, this current test makes the assumption that true vote counts follow the integer Benford distribution. Demonstrably, they do not, as this section shows.

Recall that, by Lemma 2.4, when $\theta \in \mathbb{N}$ or when $\theta \to \infty$,

$$\mathbb{P}[\mathcal{D}_1(X) = d] = \log\left(\frac{d+1}{d}\right)$$

In elections, the upper bound 10^θ is the size of the electoral division in which we are counting the votes. It is unlikely that the division size will be an integer power of 10 or that the district sizes will be sufficiently large for the asymptotic results (Figure 2.6). Thus, it is unlikely that the integer Benford distribution will hold in true vote counts.

Finally, also note that we had to assume the distribution of the vote counts was Log-uniform (Figure 2.2). Vote counts are a product of electoral division size and proportion vote for the candidate. In countries with division sizes approximately following

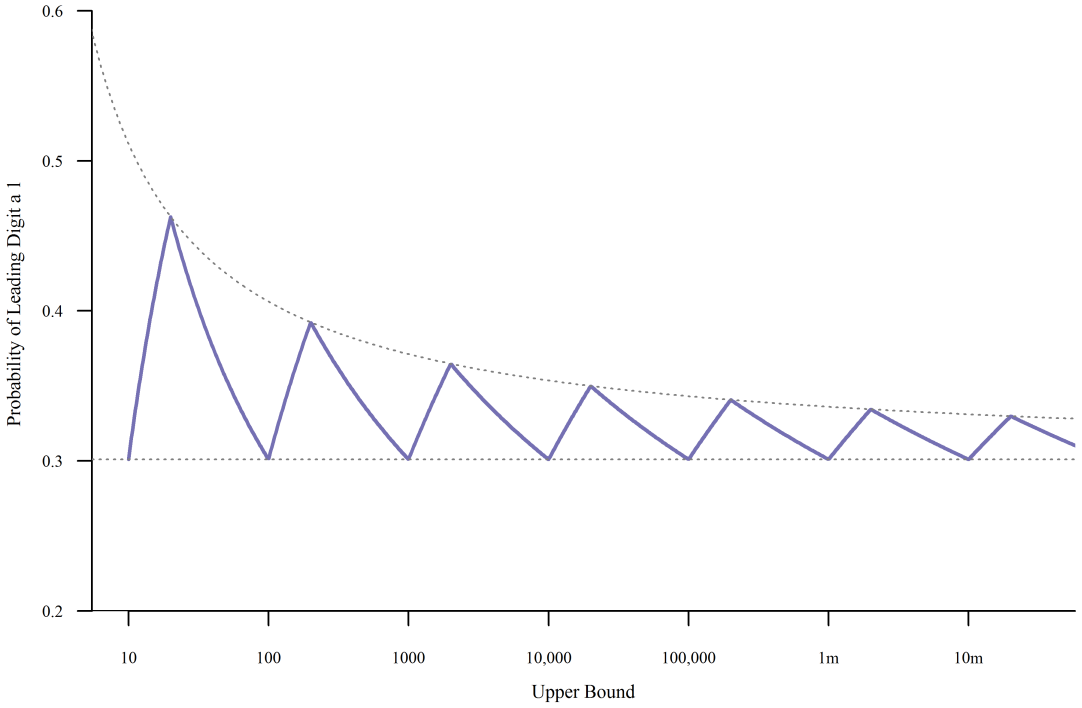


Figure 2.6: *The probability the leading digit is a '1' for various upper bounds on the Log-uniform distribution. Note that there is much variation, and the value 0.301 is the exact probability only when the upper bound is an integer power of 10.*

an Exponential distribution, such as in the United States (Figure 2.7, top panel), this requirement reduces to assuming the proportion of the vote for a given candidate is *Uniformly distributed across the electoral divisions*. Such an outcome would be quite surprising. More likely is that the distribution of proportions is unimodal, bell-shaped, and bounded between 0 and 1, much like the empirical distribution shown in Figure 2.7, bottom panel.

That raises the question of how good the Benford distribution *could* be at detecting vote-counting fraud in elections. Hill (1995) concluded that the Benford distribution is the limiting distribution of sums of random distributions. Are vote counts the sums of distributions? Yes. However, how many electoral divisions are needed for Hill's asymptotic results to hold?

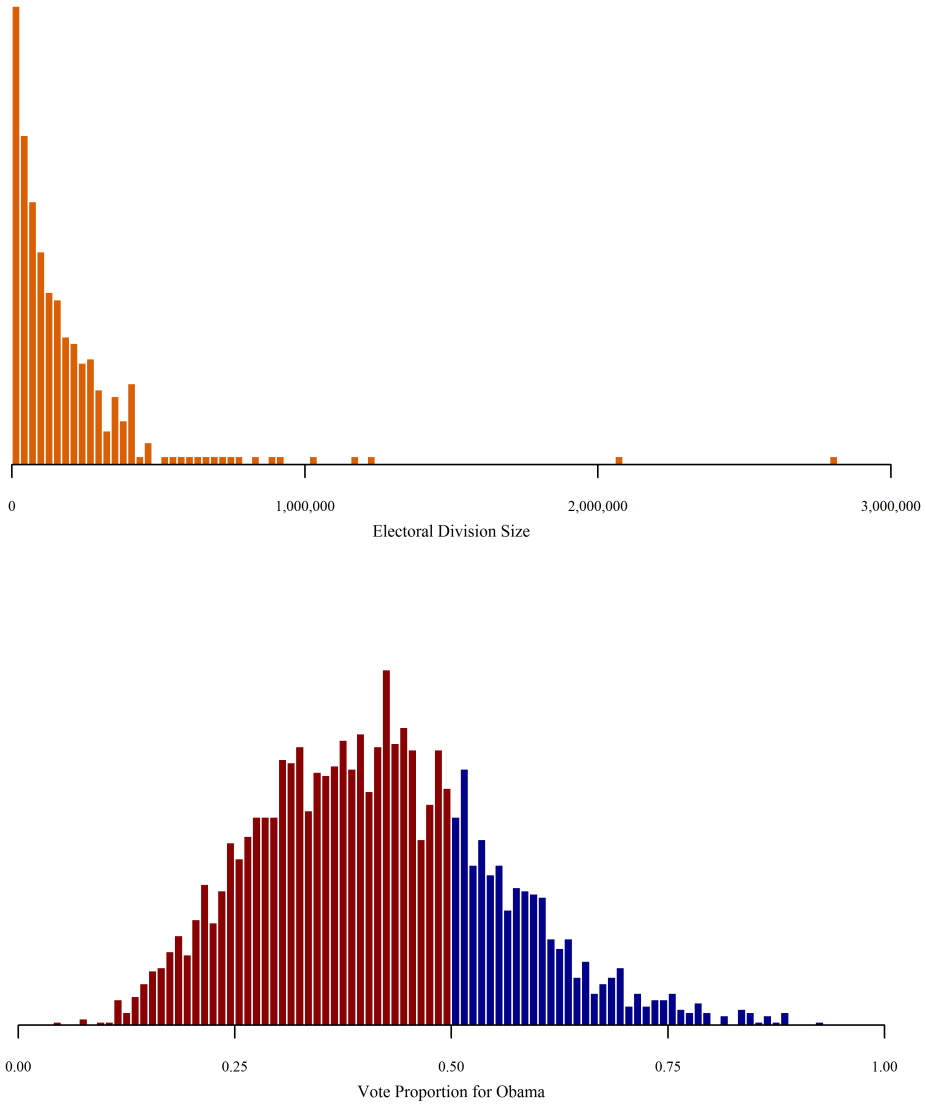


Figure 2.7: *Top panel: Distribution of electoral divisions (counties) in the United States. The vertical axis is logged frequency to show better the distribution of division sizes. Note that the distribution is similar to the Exponential distribution. Bottom panel: Distribution of proportion of vote in support of Barack Obama in the 2008 US presidential election, by electoral division. Note that this distribution is not Uniform.*

2.3.2 Weaknesses. To test if the integer Benford test is appropriate for election counts, let us use the results from the 2008 US presidential election. This datafile contains the final vote counts at the county level for every state except for Alaska ($n = 3114$ divisions). The number of counties per state ranges from 3 for Delaware to 254 for Texas. The total number of cast ballots in each county ranges from 79 in Loving County, Texas,

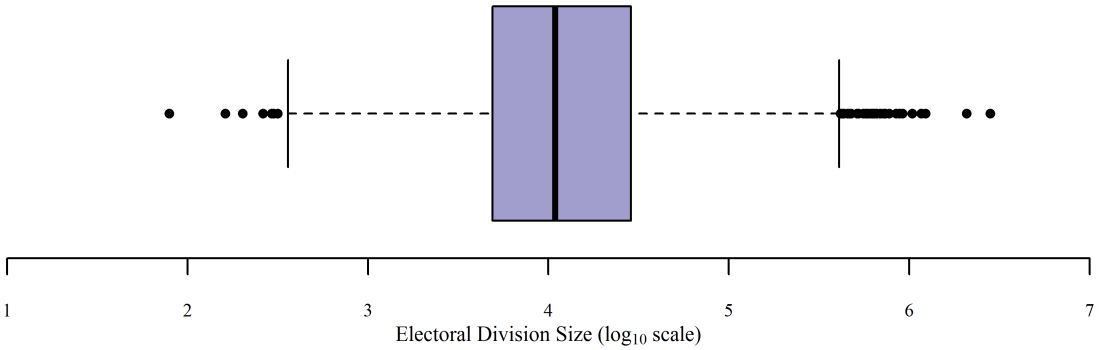


Figure 2.8: *A box-and-whiskers plot of electoral division sizes in the 2008 US presidential election. Solid dots represent outliers using the 1.5 rule.*

to 2,818,964 in Los Angeles County, California. Figure 2.8 is a box-and-whiskers plot of the distribution of electoral district sizes. Note that the horizontal scale is logarithmic.

To illustrate the appropriateness of the Benford test, let us now determine if the United States, as a whole, passes the test with the number of votes for John McCain. First, here is a table of the expected and observed proportions of leading digits at the division level.

Digit	1	2	3	4	5	6	7	8	9
Expected	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046
Observed	0.290	0.169	0.127	0.105	0.083	0.070	0.059	0.050	0.047

Now, using Pearson's Chi-squared test (Pearson 1900), we cannot reject the null hypothesis that the data follows the Benford distribution ($\chi^2 = 5.8188, \nu = 8, p = 0.6675$). Additionally, the mean digit is 3.497, which is well within the simulated 95% confidence interval (3.354, 3.527) as well as the symmetric 95% confidence interval derived from the Normal approximation, $(3.354, 3.526)$.

In the United States, election laws and procedures vary at the state level; the presidential election is actually 51 separate elections. Thus, a better assessment of this test is to examine *each state* for a violation of the integer Benford law. Unfortunately, the number of counties in several states makes the Chi-squared test inappropriate due

to the small expected cell counts (Conover 1999; Pearson 1900; Rice 2007; Yates 1934). Focusing on states whose cell counts all exceed 3 and performing the integer Benford test on the remaining 23 states, we have three states violating the Benford test at the $\alpha = 0.05$ level: Illinois, Iowa, and Mississippi. Were this test a level-α test, we would expect 1.15 states violating.

The 2004 US presidential election has even more states violating the Benford test. Of the 28 states with a sufficient number of counties, seven violated the Benford test at the $\alpha = 0.05$ level: Connecticut, Illinois, Iowa, Maine, Mississippi, Texas, and Vermont. Were this a level-α test, we would expect only 1.4 states violating.

Between these two elections, we have 10 violations in 51 tests. Under the Binomial assumptions, such an outcome would be highly unlikely, happening in fewer than 1 in 5000 cases. Thus, if we *do* assume the two US Presidential elections were free from vote-counting fraud, we have *prima facie* evidence that this is not an appropriate test of electoral fairness.

2.4. Improvements to the Benford Test

There are two major weaknesses to the integer Benford Test. The first weakness is that it assumes the upper bound of the underlying Uniform distribution (the division size) is either an integer power of 10 or is large enough for us to rely on asymptotic results. Unfortunately, the first is unlikely, and the second is not true for reasonable division sizes (*cf.* Figures 2.5 and 2.6, where there are still large variations in probabilities even when the division size is one million).

The second weakness is that the Benford test assumes the vote count follows a specific distribution—the Log-uniform distribution. There is little research regarding the distribution of votes in a free and fair election. Deckert et al. (2011) create a hierarchical probability model to simulate fraud-free election counts. They then used these elections to determine the suitability of the integer Benford test. However, even they acknowledge

that "there is no proscribed [*sic*] model of an election with which to begin the generation of artificial data" (Deckert et al. 2011, p 249).

Thus, improving upon the Benford test in this setting requires avoiding these two weaknesses. Generalizing the integer Benford test to incorporate the turnout in each division separately avoids the first weakness; removing or altering the current distributional assumption, the second.

2.4.1 The Generalized Benford Test. The integer Benford test assumes that the size of the electoral division either is an integer power of ten or is sufficiently large for asymptotic results to hold. Neither assumption reflects reality. Thus, a first correction would be to apply Theorem 2.7 separately to each division. Tests that center on this adjustment I term "generalized Benford tests," of which the integer Benford test is a special case.

Note that the above paragraph introduces one difficulty: How do we combine these n different leading digits with n different distributions into a single test statistic with a known—or even knowable—distribution?

I suggest two methods for implementing the class of generalized Benford tests: the likelihood simulation method and the multinomial averaging method. Assessment of these tests will wait until Section 2.5

The Likelihood Simulation Method: One method, albeit computationally intensive, to unify these n tests is to create a test statistic and determine its distribution through simulation. This I propose here, using the probability of observing the individual leading digits in each electoral division as that test statistic.

Let us define θ_i such that 10^{θ_i} is the total number of votes cast in electoral division i. Let c_i be the vote count in electoral division i for a specific candidate, and let $d_i := \mathcal{D}_1(c_i)$ be the leading digit of this vote count. If the leading digit follows the generalized Benford distribution, then the probability mass function in division i is given by Theorem

2.7, with θ indexed by the electoral division, i. The likelihood of observing these data is the product of the individual probabilities, with the log-likelihood being its natural logarithm:

$$\ell(\boldsymbol{\theta}; \mathbf{d}) = \sum_{i=1}^{n} \ln \left(\mathbb{P}\left[\mathcal{D}_1(X) = d_i \mid \theta_i \right] \right)$$

To estimate the distribution of the log-likelihood, one generates a leading digit in each division according to the generalized Benford distribution. The probabilities of each of those generated leading digits are multiplied together to create the likelihood value for a single election. Performing this multiple times allows one to estimate $(1 - \alpha)100\%$ confidence intervals for the likelihood value.

As a first example, let us turn to the 2008 US Presidential election in Oklahoma. The vote counts are aggregated to the county level ($n = 77$). The observed log-likelihood is -65.3186. Using simulation with 10,000 replications, an estimated 95% confidence interval is from -70 to -59. Thus, by this test, there is no evidence of vote count fraud in the 2008 US Presidential election in Oklahoma—as expected.

As a second example, let us examine the 2009 Afghan Presidential election. Recall that these votes are aggregated at the *wilāyat* level, $n = 34$. The observed log-likelihood is -28.93255. An estimated 95% confidence interval is from -32 to -25. Thus, this test also offers no evidence of vote count fraud in this election.

Note again that this section merely offers a proof-of-concept for the test. It does not offer evidence of its applicability beyond the fact that the test seems reasonable. In a later section, I explore the Type I and the Type II Error rates (Section 2.5) of this test.

Multinomial Averaging: In lieu of using simulation to estimate the distribution of a test statistic, we can estimate the test statistic in another way. Recall that the leading digit in each electoral division has a different distribution, which is a function of θ_i. That these θ_i are usually unique makes creating the correct distribution quite difficult.

Ramachandran (1956) introduced the Union-Intersection method for cases in which the null hypothesis is the intersection of several independent component null hypotheses (Berger and Sinclair 1984; Casella and Berger 2002). This method reduces the problem of testing the composite null hypothesis to testing each individual component hypothesis at an appropriately selected level. Unfortunately, unless two or more electoral divisions share a vote total, the rejection region in each division is empty; that is, the symmetric 95% confidence interval includes all nine digits. As such, the Union-Intersection method is not helpful in this situation.

However, two simple methods present themselves in terms of combining the n electoral division distributions into one.

Note that the distribution of digits in each division is a Multinomial distribution, with each distribution being a function of θ_i. That is, define $\boldsymbol{\pi}_i$ as the vector of digit probabilities in division i and $\boldsymbol{\pi}^*$ as the vector of digit probabilities in the entire country. If

$$\mathcal{D}_1(X_i) \overset{\text{iid}}{\sim} \mathfrak{B}_1(0, \theta_i),$$

then we know

$$\mathcal{D}_1(\mathbf{X}) \sim \mathcal{M}ulti\,(1, \boldsymbol{\pi}^*)\,.$$

The remaining problem is to determine the vector $\boldsymbol{\pi}^*$. Note that $\mathcal{D}_1(\cdot)$ is the leading digit function (Definition 2.2).

First: Perhaps the simplest method is to average the θ_i and use this value to calculate the $\boldsymbol{\pi}^*$ values from the Benford distribution. The main advantage to this method is that the calculation is performed easily. The main disadvantage is that there is no guarantee the *true* digit distribution follows a Benford distribution; this method forces it. Regardless, I test its performance later.

Second: Another simple method to estimate $\boldsymbol{\pi}^*$ is to calculate the mean of the digit distributions in the n divisions. Two advantages of this estimation procedure are that the calculation is simple and that the resulting digit distribution is *not* constrained

to be a member of the Benford family. The drawback is that the arithmetic average will only be an estimate of the true $\boldsymbol{\pi}^*$. However, because of its advantages, I also test its performance.

NINE-DIMENSIONAL TESTS: The previous subsection covered two methods of producing a hypothesized null distribution with the understanding that the data would be compared to it. Unfortunately this is an eight-dimensional problem in that we are projecting the nine-dimensional observation space onto a one-dimensional testing space. The default methods seem to be Pearson's Chi-squared test (Pearson 1900) and the likelihood ratio test (Neyman and Pearson 1933). However, are these the best methods in this setting?

Cressie and Read (1984) asserted that Pearson's Chi-squared test is a special case of the larger family of power-divergence tests, of which the likelihood ratio test is also a member. Each member of the power-divergence family of statistics has the form of

$$\frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{k} x_i\left[\left(\frac{x_i}{\mu_i}\right)^{\lambda}-1\right]$$

where $\lambda \in \mathbb{R}$ is a parameter selecting the specific form of the test, x_i is the ith observed value, and μ_i is the ith expected value. When $\lambda = 1$, we have Pearson's Chi-squared statistic, X^2. When $\lambda \to 0$, we have the usual likelihood ratio statistic, G^2. Cressie and Read (1984) suggest that the statistic with $\lambda = \frac{2}{3}$ tends to be superior to either of the above two options. All members of the power-divergence family have the same limiting Chi-squared distribution (Read and Cressie 1988, p 47).

Such tests are not the only way of projecting nine dimensions into one. Another method is to focus on the p-values from the individual nine tests and select just the minimum, the min-p. However, due to multiple testing issues, one would have to adjust this value in some manner. The default method is the Bonferroni adjustment where the calculated p-values are multiplied by the number of tests performed, k (Holm 1979). While this does protect the experiment-wise rejection rate, it is conservative—*especially*

when the p-values are discrete random variables, as is the case here (Murdoch et al. 2008; Westfall and Wolfinger 1997).

There is an adjustment procedure that takes into consideration the discreteness of the p-values, creating a higher power than the Bonferroni adjustment, while still protecting the experiment-wise error rate. The Westfall-Wolfinger min-p test (a Union-Intersection test) defines the adjusted p-value as

$$p'_j := \mathbb{P}\Big[\min_i \{P_i\} \leq p_j\Big] \tag{2.1}$$

Here, the P_i are the p-values (random variables) of the k tests, and p_j is the smallest observed p-value, which corresponds to that of the j^{th} test.

Following Westfall and Wolfinger (1997), let us further define p_i as the observ*ed* p-value of the i^{th} test and p_{it} as the observ*able* p-values of the i^{th} test. With this, and with the definition of the p-value, we have $p_{it} = \mathbb{P}[P_i \leq p_{it}]$. Since the p-value is defined as the probability under the null hypothesis of observing data this extreme (or more so), we can define

$$p_{it(j)} := \begin{cases} \max_t \{p_{it},\ \text{s.t. } p_{it} \leq p_j\} & \text{if } \min_t \{p_{it}\} \leq p_j \\ 0 & \text{Otherwise} \end{cases} \tag{2.2}$$

as the p-value of each of the i^{th} tests, as a function of the minimum observed p-value (in digit j). With this, the adjusted p-value is

$$p'_j = 1 - \prod_{i=1}^{k} \Big(1 - p_{it(j)}\Big)$$

As this holds for all observed p-values, it also holds for the minimum of the observed p-values. And so, to create the most powerful test of this class, one need only calculate p'_j for the digit with the smallest p-value.

For the current situation of digit tests on n electoral divisions, both i and j range from 1 to 9, and t ranges from 0 to n. The annex to this chapter has the `R`-code for this test (see page 59).

Freq. (t)	Digit 1 p-value	Digit 2 p-value	Digit 3 p-value	Digit 4 p-value	Digit 5 p-value	Digit 6 p-value	Digit 7 p-value	Digit 8 p-value	Digit 9 p-value
0	0.559	0.641*	1.000*	1.000*	1.000*	1.000*	1.000*	1.000	1.000*
1	1.000*	1.000	0.330	0.263	0.219	0.188	0.164	0.146	0.131
2	0.217	0.082	0.043	0.026	0.018	0.013	0.010	0.008*	0.006
3	0.027	0.005	0.002	0.001	0.000	0.000	0.000	0.000	0.000

Table 2.2: *Table of observable p-values for the example in the text. Starred p-values are observed. The red p-value is the minimum of the observed p-values.*

To illustrate this test, let us examine a toy example in which there are only three electoral divisions. The vote counts in the divisions are 1405, 8037, and 8204. Using reflected tail probabilities for a two-tailed test (Westfall and Wolfinger 1997), we have the observable p-values given in Table 2.2. Starred values are observed p-values given the observed data.

According to Table 2.2, we have:

$$\begin{aligned} i &= 1, 2, \ldots, 9 \\ j &= 8 \\ t &= 0, 1, 2, 3 \\ p_j &= 0.008 \\ p_{it(j)} &= \big(0.000, 0.005, 0.002,\ \ 0.001, 0.000, 0.000,\ \ 0.000, 0.008, 0.006\big)', \text{ and} \\ p'_j &= 0.02182363 \end{aligned}$$

Note that this adjusted p-value is less than the one would calculate using the Bonferroni adjustment ($p = 0.008 \times 9 = 0.072$). In fact, one would fail to reject the null hypothesis that the leading digit follows the Benford distribution when using the Benford test. One would reject the null hypothesis using the Westfall and Wolfinger min-p test.

And so, we are left with the question of which test is most appropriate and when. To determine this, I test the level and power of these four tests: Pearson's Chi-squared test ($\lambda = 1$), the likelihood ratio test ($\lambda \to 0$), Read and Cressie's suggested test ($\lambda = \frac{2}{3}$), and Westfall and Wolfinger's min-p test. The results of the level tests (Figure 2.9) suggest that two of the four tests are conservative. The Read and Cressie test is only slightly conservative. The likelihood ratio test is highly conservative when the number of divisions

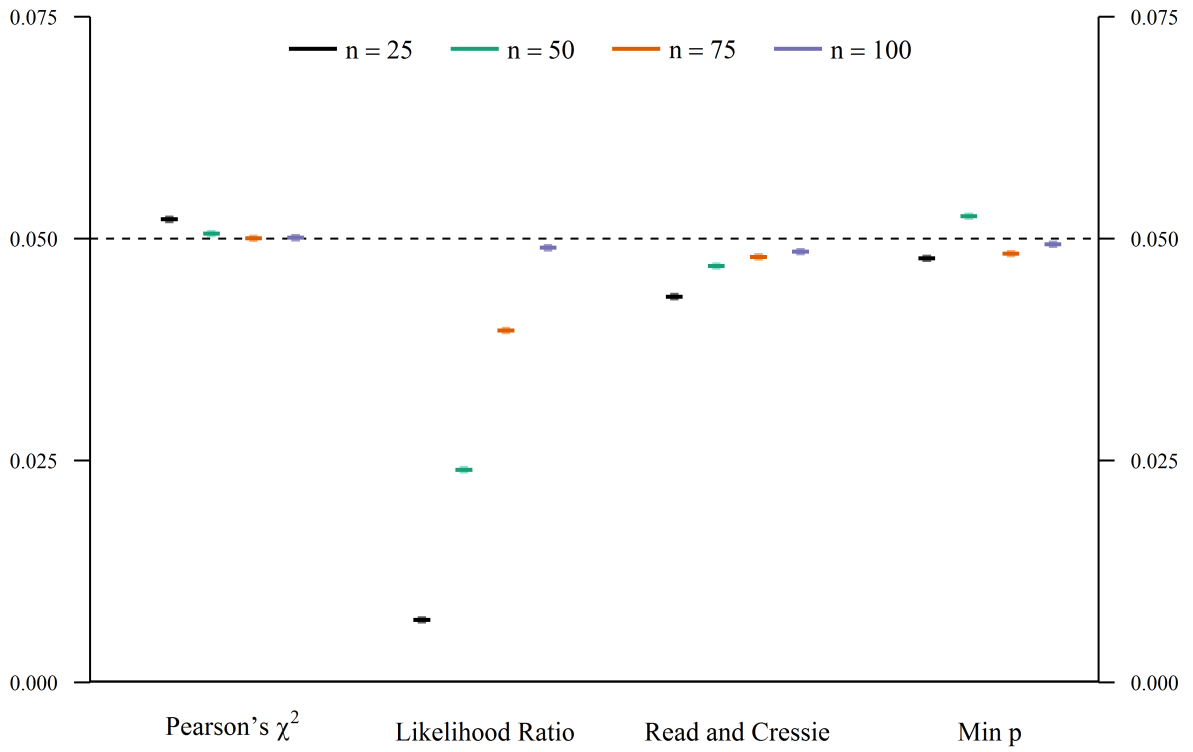


Figure 2.9: *A plot of the estimated sizes of the four tests when* $\alpha = 0.05$ *for four sample sizes. In each, the 95% confidence intervals are shown as rectangles. The number of iterations to estimate the Type I Error rate is* $B = 1,000,000$*. Note the issue the likelihood ratio test,* G^2*, has when the digit frequencies in bins are small.*

is low, reducing as the number of divisions increases. Both the Chi-squared test and the min-p test are very close to being level-α tests.

To test power, I specify two alternative distributions: the discrete Uniform distribution and a shifted Benford distribution. The first alternative distribution would arise from people fabricating the vote counts; the second, from adding one to the leading digit.

With these two alternative distributions, the power-testing program parameters are

- The null distribution is $H : D \sim \mathfrak{B}_1(0, n)$, with $n \in \mathbb{N}$.
- The alternative distributions are

$$K_u : D \sim \mathcal{U}(1, 9),$$
$$K_s : D \sim \mathfrak{B}_1^s(0, n).$$

- The mixing parameter, ξ varies from 0 to 1 in increments of 0.10.
- The sample size, n, varies from 25 to 100 in increments of 25.
- The four tests are performed and the rejection rate is calculated for each at the $\alpha = 0.05$ level based on $B = 100,000$ trials.

Running this in `R` would have taken days. Thus, I wrote the code in `FORTRAN 90` and enjoyed a nice cup of Kenyan coffee while it ran.

Figure 2.10 displays the estimated power curves for each of the tests as functions of the mixing parameter, with the sample size held at 75. One thing is notable here: the venerable Chi-squared test is superior to the other tests for detecting uniform contamination; the Westfall and Wolfinger min-p test, for detecting Shift contamination. Both will need to be used.

2.4.2 The Empirical Benford Tests. The previous methods still assume that the vote counts have a specific distribution—the Log-uniform distribution. If we are not comfortable assuming this distribution for vote counts (Deckert et al. 2011), we have options. First, we could change the assumed distribution. Second, we could eliminate the distributional assumption completely and use non-parametric methods.

The Parametric Empirical Benford Test: The parametric empirical Benford test generates multiple "elections" based on the distribution of the proportion of the vote in favor of a specified candidate and calculates the distribution of leading digits according to this electoral system. In one manner, this method is similar to the likelihood simulation method for the generalized Benford test (Section 2.4.1); simulation is used to determine an important distribution. Here, however, that important distribution is the distribution of leading digits. The difficult parts comes down to generating elections from the given data.

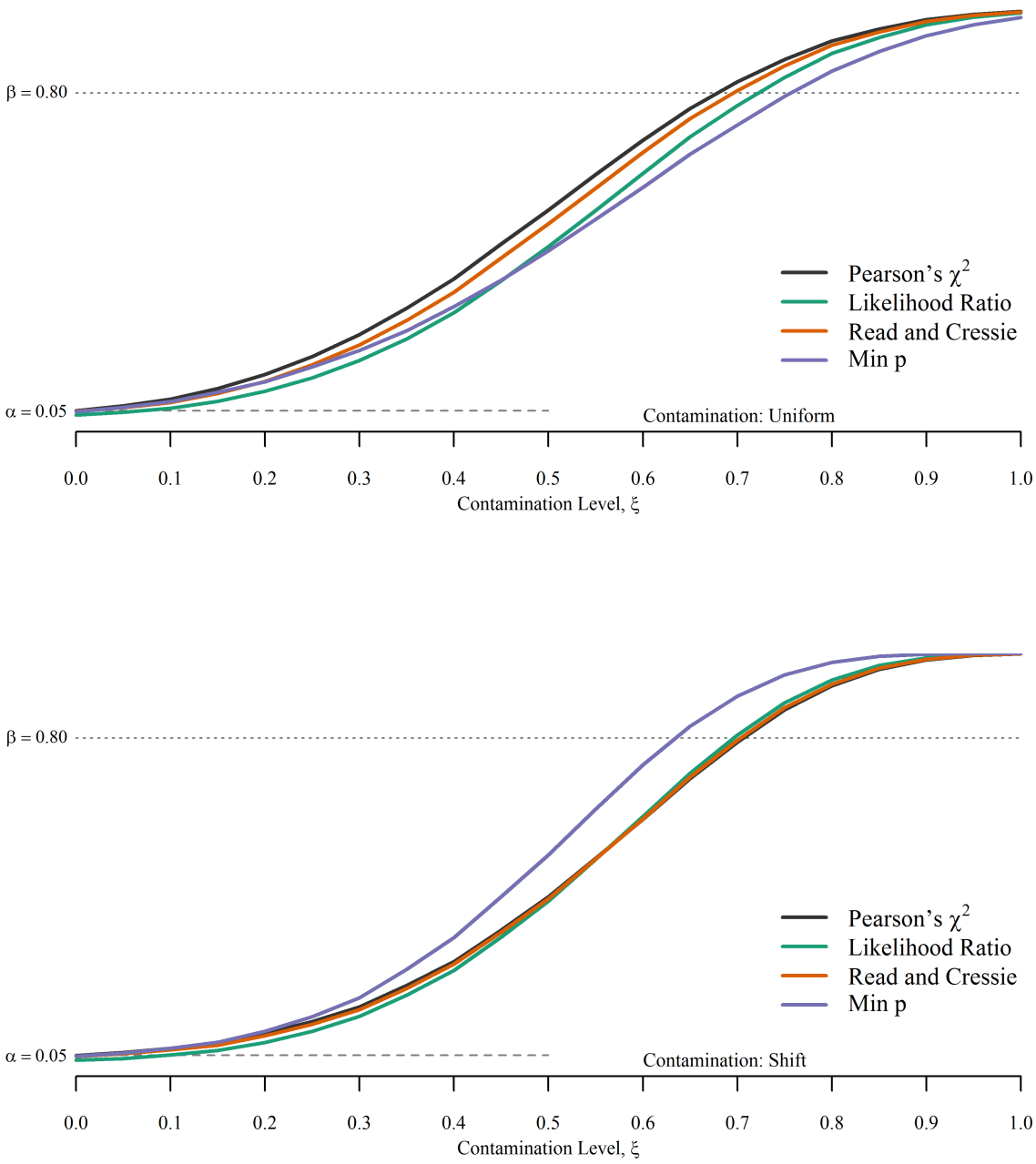


Figure 2.10: *A plot of the estimated power curves for each of the four tests when* $\alpha = 0.05$ *and* $n = 75$*. The number of Monte Carlo iterations is* $B = 10,000$*. This corresponds to an estimated margin of error of* ± 0.01*.*

Let us recall the bottom panel of Figure 2.7 on page 25. As in this case, the distribution of vote proportions for a specified candidate often follow a unimodal distribution with support in $(0, 1)$. While there are several distributions matching these requirements,

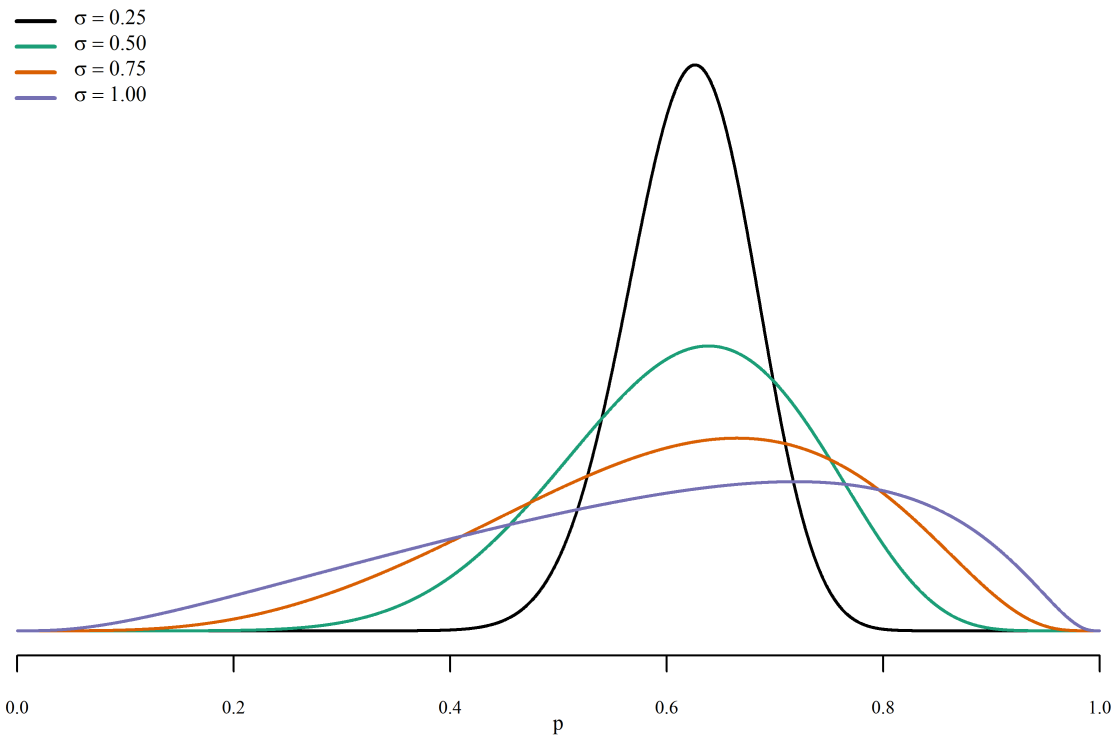


Figure 2.11: *Plots of the Logit-normal* $\text{Lgt}\mathcal{N}\left(\mu, \sigma^2\right)$ *distribution for* $\mu = 0.60$ *and for four values of* σ*.*

one helpful distribution is the Logit-normal distribution (Johnson 1949).

Definition 2.9 (Logit-normal distribution)**.** *Let* $Y \sim \mathcal{N}\left(\mu, \sigma\right)$*. We say* $X := \text{logit}\, Y$ *has a Logit-normal distribution, symbolized as* $X \sim \text{Lgt}\mathcal{N}\left(\mu, \sigma^2\right)$*, with support* $X \in (0, 1)$*.*

The probability density function of the Logit-normal distribution is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x(1-x)} \exp\left[-\frac{(\text{logit}\, x - \mu)^2}{2\sigma^2}\right]$$

Its cumulative distribution function is

$$F(x) = \Phi\left(\frac{\text{logit}\, x - \mu}{\sigma}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard Normal distribution. Figure 2.11 provides plots of Logit-normal distributions for four values of σ with $\mu = 0.60$.

Alas, the moments of the Logit-normal distribution lack analytic solutions unless $\mu = 0.500$. Regardless, calculations of these moments are straightforward using numerical methods and statistical packages (Frederic and Lad 2008; Johnson 1949; Wutzler 2012). An advantage of the Logit-normal distribution over the more common Beta distribution is that a logit transformation of the Logit-normal distribution produces a Normal distribution, and the Normal distribution frequently has better properties (both mathematical and statistical) than the Beta distribution.

In terms of the parametric empirical Benford test, this distribution serves as the generator distribution for the proportion of vote in favor of the candidate in each electoral division. The process of generating elections is typical for simulation studies:

1. The μ and σ^2 parameters are calculated from the data;

2. a random draw of length n from a Logit-normal $Lgt\mathcal{N}(\mu, \sigma^2)$ distribution is produced;

3. element-wise multiplying this vector by the vector of division sizes produces vote counts produces a pseudo-election; finally

4. from this pseudo-election, a leading digit distribution is calculated.

Repeating this process increases the precision of the leading digit distribution. Comparing the observed leading digit distribution to this empirically-generated distribution is the test. The comparison test should be the min-p test or Pearson's Chi-square test (see §2.4.1).

For instance, let us return to the 2008 US presidential election. Figure 2.12 provides a histogram of the vote proportion for candidate John McCain in all US counties, overlain by the Logit-normal $Lgt\mathcal{N}(\mu = 0.298, \sigma = 0.616)$ distribution.

In this election, a total of 3114 electoral divisions reported vote counts, with division sizes ranging from 79 (Loving County, TX) to $2,819,000$ (Los Angeles County,

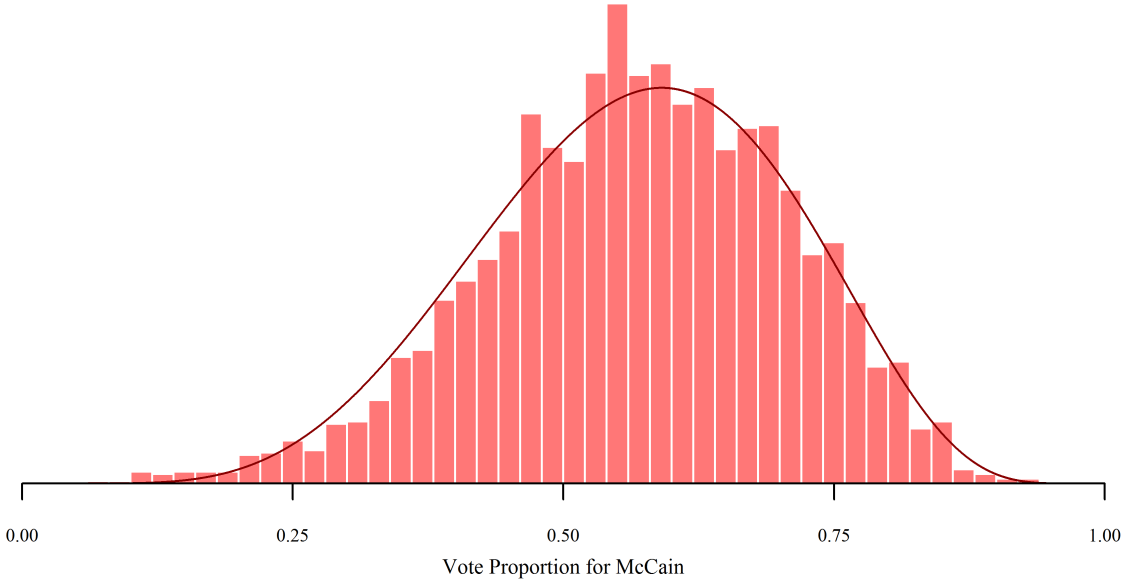


Figure 2.12: *Histogram of McCain support by county in the United States overlaid with the Logit-normal* $Lgt\mathcal{N}(\mu = 0.298, \sigma = 0.616)$ *distribution.*

CA). In this election, electoral divisions tended to be counties; however, some states reported at other levels, such as cities and townships. With $B = 10,000$ replications, the table below provides the expected distribution and observed proportions of leading digits.

Digit	1	2	3	4	5	6	7	8	9
Expected	0.286	0.175	0.129	0.102	0.084	0.070	0.059	0.051	0.045
Observed	0.290	0.169	0.127	0.105	0.083	0.070	0.059	0.050	0.047

Figure 2.13 displays the expected frequency of the leading digits as well as the observed frequencies. The Pearson Chi-squared test indicates that the observed digit distribution does not significantly deviate from the expected digit distribution ($X2 = 1.43; p = 0.994$). This conclusion is echoed by the generalized likelihood test ($G2 = 0.716; p = 0.999$) and Read & Cressie's test ($RC = 1.43; p = 0.994$). The min-p test also concurs ($p' = 0.994$). In all four cases, there is no sufficient evidence for vote-counting fraud.

As a second example, let us revisit the Afghan 2009 Presidential election. Recall that a total of 34 *wilāyet* reported vote counts in this election, with division sizes

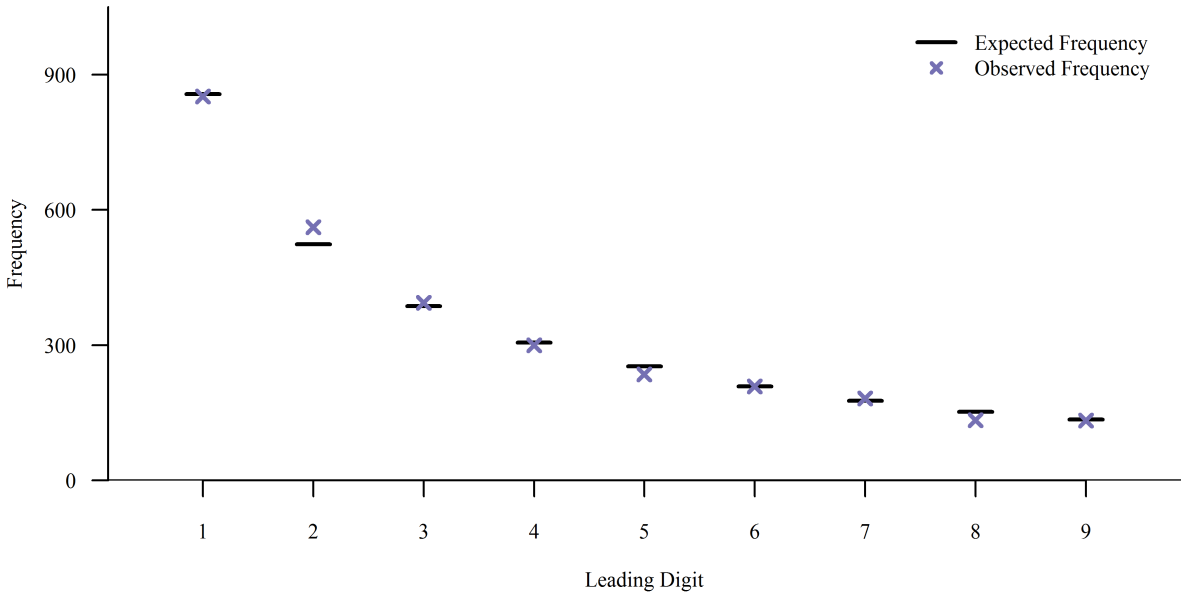


Figure 2.13: *Graphic of expected versus observed leading digit frequencies for John McCain's votes in the 2008 US Presidential election using the parametric empirical Benford test.*

ranging from $11,300$ to $152,800$. Using $10,000$ replications, the expected and observed distribution of leading digits is given in the table below.

Digit	1	2	3	4	5	6	7	8	9
Expected	0.295	0.182	0.120	0.104	0.084	0.069	0.057	0.049	0.042
Observed	0.265	0.235	0.059	0.147	0.118	0.059	0.088	0.029	0.000

Figure 2.14 displays the expected frequency of the leading digits as well as the observed frequencies. The Pearson Chi-squared test indicates that the observed digit distribution does not significantly deviate from the expected digit distribution ($X2 = 5.09; p = 0.747$). Because of the zero count, the generalized likelihood statistic does not exist. The Read and Cressie test concurs with the conclusion of the Chi-squared test ($RC = 5.37; p = 0.717$). The min-p test also suggests no significant evidence of vote-count fraud ($p' = 0.905$).

The advantage of the parametric empirical Benford test is that it removes the assumption that the vote counts have a Log-uniform distribution. There are two disadvantages. The first is that this test is of low power (see Section 2.5). The second is that

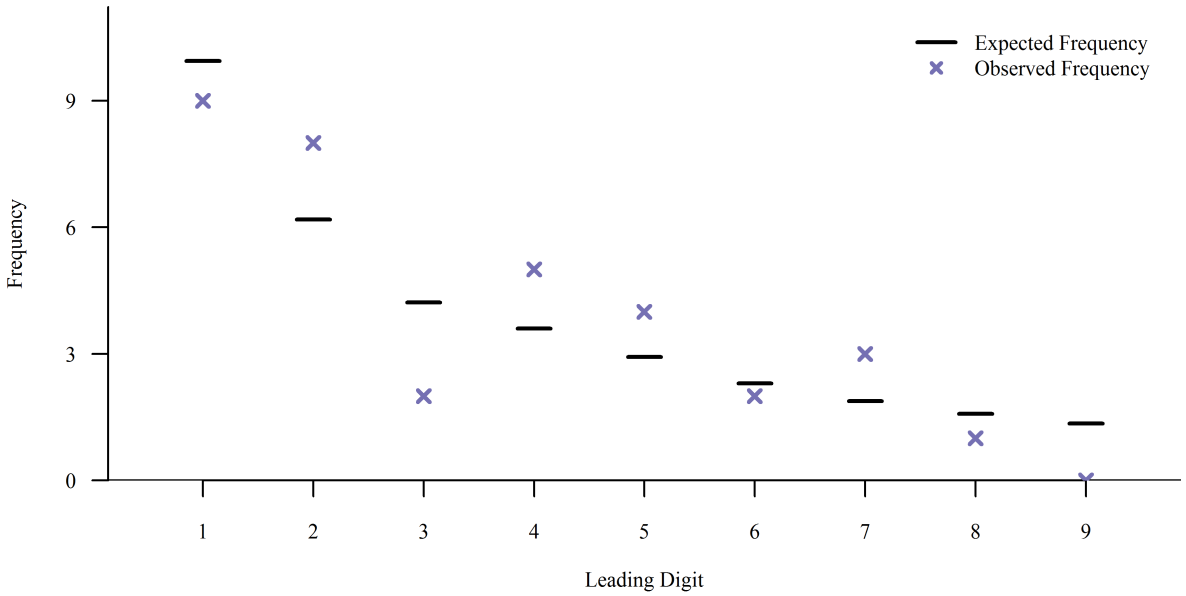


Figure 2.14: *Graphic of expected versus observed leading digit frequencies for Hamid Karzai's votes in the 2009 Afghan Presidential election using the parametric empirical Benford test.*

it makes the assumption that the vote proportions follow the Logit-normal distribution. There is little extant research regarding the *a priori* distribution of vote proportions. To adjust the test for this concern, I next relax this last assumption.

The Non-Parametric Empirical Benford Test: Non-parametric tests are used when the distribution of the data is not known, such as here. The non-parametric empirical Benford test is very similar to the parametric empirical Benford test in that the expected distribution of leading digits is generated through simulation. The primary difference is that the simulated vote proportions are randomly drawn from the real vote proportions rather than from the Logit-normal distribution. Thus, the steps are

1. Generate a random sample of length n from the vote proportions, with replacement;
2. create vote counts for a pseudo-election by element-wise multiplying the vote-proportion vector by the division-size vector; and
3. calculate the leading-digit distribution from this pseudo-election.

As before, repeating this process increases the precision of the empirical leading digit distribution, and the statistical test can be done using the min-p test or a member of the power-deviance family of tests (Section 2.4.1, above).

2.5. The Alpha and the Beta

In this chapter, I have covered several current and proposed digit tests. Which of these tests, if any, should be used and under which circumstances? This question can be answered by looking at the level and power of each test.

For a test to be of level α, the rejection rate *under the null hypothesis* should not exceed α. Investigating the level of the test requires that we know the distribution of a free and fair election. Unfortunately, Political Science has not advanced to the point that there is a known distribution for vote counts (Deckert et al. 2011). In lieu of using a hypothesized distribution of a free and fair election, I shall use the vote counts generated from an election I *assume* is fair. This election is the US Presidential election of 2008—the full complement of the 50 states excluding those reporting fewer than five electoral divisions in the election. This collection offers divisions of several sizes and elections with different numbers of divisions.

As discussed in Section 1.3, I extract the division size and the proportion of the vote in that division cast for John McCain—the candidate of the incumbent party. I then permute the proportions and multiply them by the division sizes. Repeating this created multiple elections. The R code is provided in the Annex 2 (Section 2.8).

With these assumedly free and fair elections, we can estimate the size of each of the tests of this chapter. In the next two sections, I demonstrate this with the likelihood simulation and the multinomial averaging methods by investigating their level.

In terms of estimating the power of the tests, not only is a "null" distribution required (the distribution of vote counts from a free and fair election), but so is an "alternative" distribution—vote counts from an election demonstrating a lack of fairness.

The null distribution will be as above. The alternative distributions will either be the uniform contamination or the spike contamination (Section 2.4.1, page 34), depending on the test's power.

2.5.1 The Likelihood Simulation Method. Recall from Section 2.4.1 (page 28), the Likelihood Simulation method calculates the likelihood of observing the reported leading digits and compares that likelihood to the distribution of likelihoods generated from simulation of the data. The likelihoods are based on the generalized Benford distribution. The algorithm for a single test is

1. Determine the division size, 10_i^{θ}, for each division;

2. Generate an 'election' of digits based on θ_i for each division using the generalized Benford distribution; and

3. Calculate the log-likelihood of this 'election.'

Repeating these steps a sufficient number of times allows one to more precisely estimate the distribution of the log-likelihood, against which one compares the observed log-likelihood. Performing this test for the 47 elections of 2004 and the 47 elections of 2008 above gives the results mapped in Figure 2.15. Tables 2.3 and 2.4 provide the results for each of the 94 elections.

For these 94 elections, 10 of the observed likelihoods fall outside the 95% confidence interval. This corresponds to an observed rejection rate (at $\alpha = 0.05$) of 0.1064, which is more than twice the expected rejection rate. Assuming the tests are independent, which would imply the number of rejections would be Binomially distributed, there should be no more than nine rejections, corresponding to a rejection rate of 0.0957. The p-value for the hypothesis that the rejection rate is actually $\alpha = 0.05$ is 0.02311. Thus, this is an improper test, rejecting at a higher rate than it should. Admittedly, given the origin of the test, it is not as poor as expected.

Election(s)	Divisions	Lower Bound	Upper Bound	Measured	
Alabama	67	-61.75	-51.93	-58.25	
Arizona	15	-15.08	-10.39	-11.97	
Arkansas	75	-68.57	-57.97	-60.45	
California	58	-53.87	-44.60	-47.44	
Colorado	64	-59.03	-49.14	-54.91	
Connecticut	169	-150.48	-134.69	-151.35	✓
Florida	67	-61.81	-52.08	-58.00	
Georgia	159	-141.77	-126.43	-135.37	
Idaho	44	-41.14	-32.93	-35.65	
Illinois	102	-92.41	-80.08	-88.18	
Indiana	92	-83.84	-72.35	-84.65	✓
Iowa	99	-89.72	-77.73	-88.74	
Kansas	105	-94.66	-81.98	-83.16	
Kentucky	120	-107.71	-94.43	-104.48	
Louisiana	64	-59.14	-49.46	-57.74	
Maine	511	-439.14	-411.04	-433.45	
Maryland	24	-23.31	-17.42	-20.04	
Massachusetts	351	-306.96	-283.74	-298.42	
Michigan	83	-75.82	-64.83	-76.26	✓
Minnesota	87	-79.18	-67.88	-74.09	
Mississippi	82	-74.98	-64.14	-77.27	✓
Missouri	115	-103.53	-90.62	-96.38	
Montana	56	-51.73	-42.31	-43.67	
Nebraska	93	-83.98	-71.85	-74.08	
Nevada	17	-16.87	-11.74	-12.13	
New Hampshire	237	-207.72	-188.17	-194.55	
New Jersey	21	-20.55	-15.09	-18.94	
New Mexico	33	-31.41	-24.34	-27.01	
New York	62	-57.23	-47.65	-47.42	✓
North Carolina	100	-90.60	-78.47	-83.82	
North Dakota	53	-49.22	-40.16	-43.92	
Ohio	88	-80.27	-68.90	-70.53	
Oklahoma	77	-70.29	-59.44	-60.05	
Oregon	36	-34.11	-26.90	-34.59	✓
Pennsylvania	67	-61.69	-51.75	-51.15	✓
Rhode Island	39	-36.90	-29.21	-33.77	
South Carolina	46	-43.15	-35.11	-40.17	
South Dakota	66	-60.59	-50.53	-54.81	
Tennessee	95	-86.26	-74.45	-79.42	
Texas	254	-224.16	-204.53	-215.89	
Utah	29	-27.76	-21.18	-25.59	
Vermont	246	-215.25	-195.29	-212.64	
Virginia	134	-120.20	-106.30	-116.56	
Washington	39	-36.95	-29.18	-31.56	
West Virginia	55	-51.03	-42.00	-47.13	
Wyoming	23	-22.33	-16.46	-19.01	
Wisconsin	72	-66.20	-56.05	-60.86	

Table 2.3: *Table of Type I Error rates for the Likelihood Simulation Method for the 2004 US Presidential election. The number of repetitions was* $B = 10,000$.

Election(s)	Divisions	Lower Bound	Upper Bound	Measured	
Alabama	67	-61.77	-51.98	-61.40	
Arizona	15	-15.08	-10.45	-12.73	
Arkansas	75	-68.51	-57.92	-62.20	
California	58	-53.87	-44.61	-51.14	
Colorado	64	-59.04	-49.13	-55.92	
Connecticut	8	-8.47	-5.22	-5.87	
Florida	67	-61.82	-52.11	-59.54	
Georgia	159	-142.01	-126.68	-139.02	
Idaho	44	-41.16	-32.96	-35.42	
Illinois	102	-92.42	-80.14	-92.65	✓
Indiana	92	-83.83	-72.30	-85.58	✓
Iowa	99	-89.67	-77.72	-86.17	
Kansas	105	-94.76	-82.12	-82.39	
Kentucky	120	-107.72	-94.37	-102.39	
Louisiana	64	-59.16	-49.52	-58.78	
Maine	16	-16.00	-11.13	-14.42	
Maryland	24	-23.30	-17.42	-19.86	
Massachusetts	14	-14.14	-9.69	-11.46	
Michigan	83	-75.84	-64.84	-74.43	
Minnesota	87	-79.13	-67.84	-75.15	
Mississippi	82	-75.01	-64.18	-76.87	✓
Missouri	115	-103.54	-90.58	-94.62	
Montana	56	-51.74	-42.36	-43.01	
Nebraska	93	-83.98	-71.86	-75.81	
Nevada	17	-16.86	-11.72	-12.14	
New Hampshire	10	-10.37	-6.57	-8.81	
New Jersey	21	-20.54	-15.07	-18.82	
New Mexico	33	-31.39	-24.34	-26.93	
New York	62	-57.25	-47.69	-48.25	
North Carolina	100	-90.71	-78.59	-84.75	
North Dakota	53	-49.25	-40.21	-46.99	
Ohio	88	-80.20	-68.82	-70.51	
Oklahoma	77	-70.32	-59.44	-59.75	
Oregon	36	-34.24	-26.97	-33.19	
Pennsylvania	67	-61.68	-51.75	-52.76	
Rhode Island	5	-5.60	-2.93	-4.71	
South Carolina	46	-43.17	-35.09	-42.18	
South Dakota	66	-60.58	-50.57	-55.90	
Tennessee	95	-86.25	-74.40	-78.52	
Texas	254	-224.01	-204.46	-213.12	
Utah	29	-27.76	-21.17	-24.58	
Vermont	14	-14.13	-9.52	-13.63	
Virginia	134	-120.24	-106.43	-115.75	
Washington	39	-36.91	-29.18	-31.23	
West Virginia	55	-51.05	-41.91	-44.93	
Wisconsin	72	-66.18	-56.04	-63.67	
Wyoming	23	-22.33	-16.47	-17.44	

Table 2.4: *Table of Type I Error rates for the Likelihood Simulation Method for the 2008 US Presidential election. The number of repetitions was* $B = 10,000$.

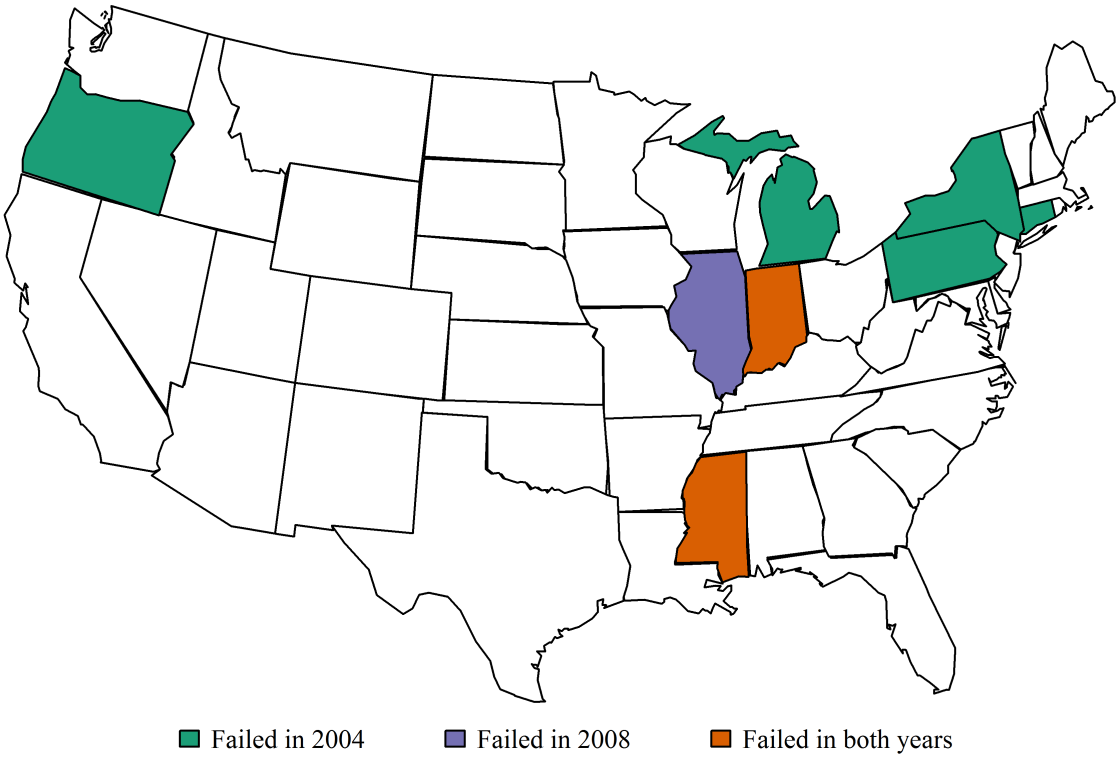


Figure 2.15: *Table of Type I Error rates for the Likelihood Simulation Method in these elections. The average rejection rate for these elections is* 0.1065.

To continue exploring the rejection rates for real elections, I performed this test on 10 additional elections. Table 2.5 provides the estimated p-value for each of those 10 elections. There were claims of electoral fraud in only the Afghan and Romanian elections. The Afghan election easily passed this test; neither Romanian election did. Unfortunately, both Irish elections and the Norwegian election soundly failed. I would say this constitutes *prima facie* evidence that the test is problematic.

I now turn to estimating the power of this test. The fully-contaminated election will be the US 2008 election with leading digits shifted. To estimate power, I create mixtures of the null election and this fully adulterated election, varying the contamination level. From this, I can estimate the rejection rate at each contamination level.

Figure 2.16 provides a plot of the calculated p-values for each election tested using this test (green dots) and the proportion of those which are below $\alpha = 0.05$ (curve). Note

State	Year	Election	LST
Afghanistan	2009	President	0.992
Egypt	2011	Referendum	0.329
Ireland	2011	Parliament	0.000
Ireland	2012	Referendum	0.000
Lithuania	2009	President	0.851
Macedonia	2011	Parliament	0.388
Norway	2009	Parliament	0.009
Romania	2009	President (Round 1)	0.000
Romania	2009	President (Round 2)	0.000
United States	2004	President	0.939

Table 2.5: *Results for the Likelihood Simulation test performed on 10 real elections. The LST column provides p-values estimated by this method. Only in the Afghan and Romanian elections was fraud alleged.*

that the contamination rates tested range from 0.00 to 0.10, not the usual 0.00 to 1.00 range.

Because the power is approximately 1.00 when the shift contamination is at 6% ($\xi = 0.06$), this test is very sensitive to the shift contamination. This may explain the Irish and Norwegian results in Table 2.5. There is little room for deviation from the assumed distribution before the test rejects the null hypothesis.

2.5.2 Multinomial Averaging. Recall from Section 2.4.1 on page 29, that I propose two methods for implementing the multinomial averaging procedure, two ways of estimating the distribution of the leading digits. The first method uses the *average division size* to construct the expected distribution of leading digits (MA1). The second method *averages the expected leading digit frequencies* across the electoral divisions (MA2). In both cases, if the division sizes are known, the final distribution does *not* change (Figure 2.17).

To test the power of these methods, I simulated elections and performed the tests, determining the rejection rate for various alternative distributions. The null elections (distributions) were the observed 2008 Presidential election results from 47 US states.

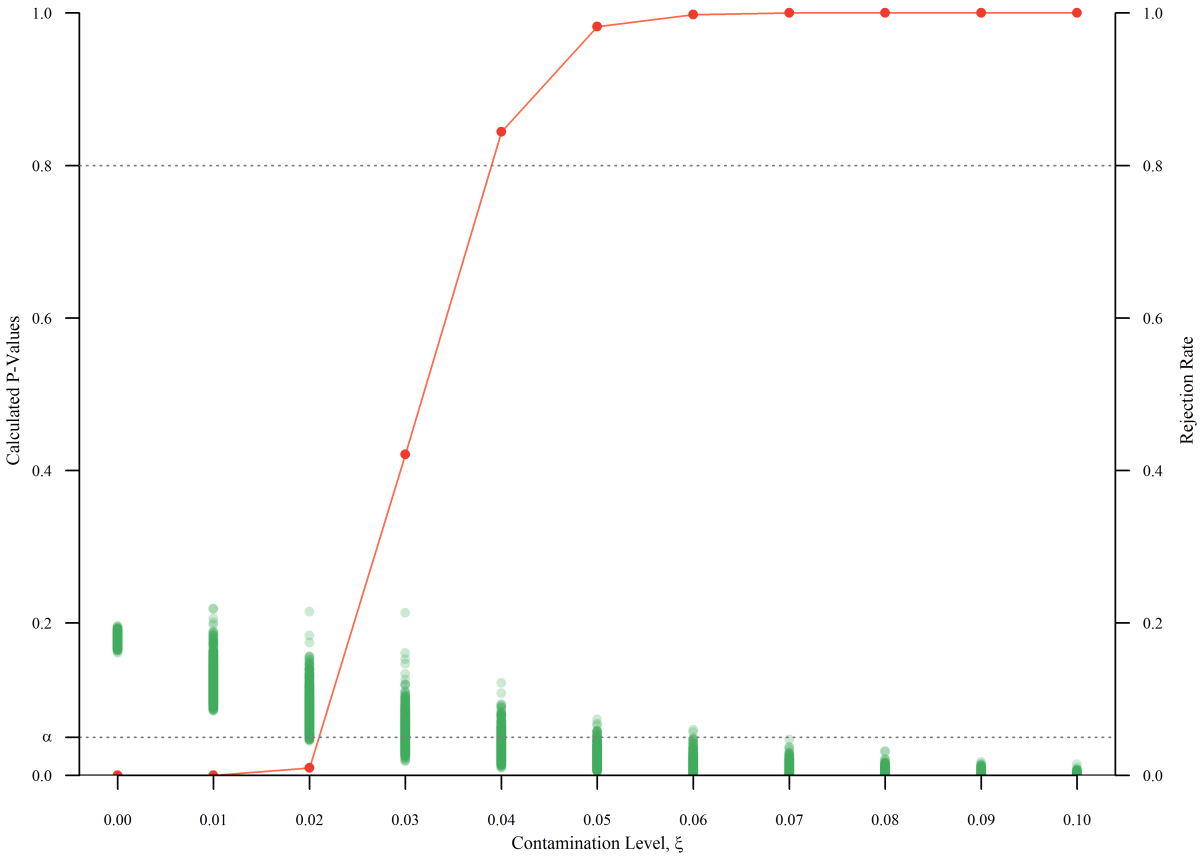


Figure 2.16: *Plot of calculated p-values and the resulting power curve for the Likelihood Simulation Method in these elections. There are 1000 elections tested at each contamination level.*

From those vote counts, I replaced the leading digit with the next higher digit, representing the *shift* contamination. These are the fully contaminated elections, where all electoral divisions are tainted.

As I am interested in the power of the test, I created nine additional sets of elections blending the null and the fully-contaminated elections, where the blending rate ranged from 0% to 15%, in steps of 1%. Thus, each of the 47 states had 10 "alternative" sets of elections. Each set of elections contained 10,000 individual elections.

I performed the two versions of the multinomial averaging procedure, MA1 and MA2, on each of the generated elections using Pearson's Chi-square test. Figure 2.17 shows the expected digit distributions for the MA1 and MA2 version. Note that the two distribution are very similar. The largest differences occur for digits 2 through 4.

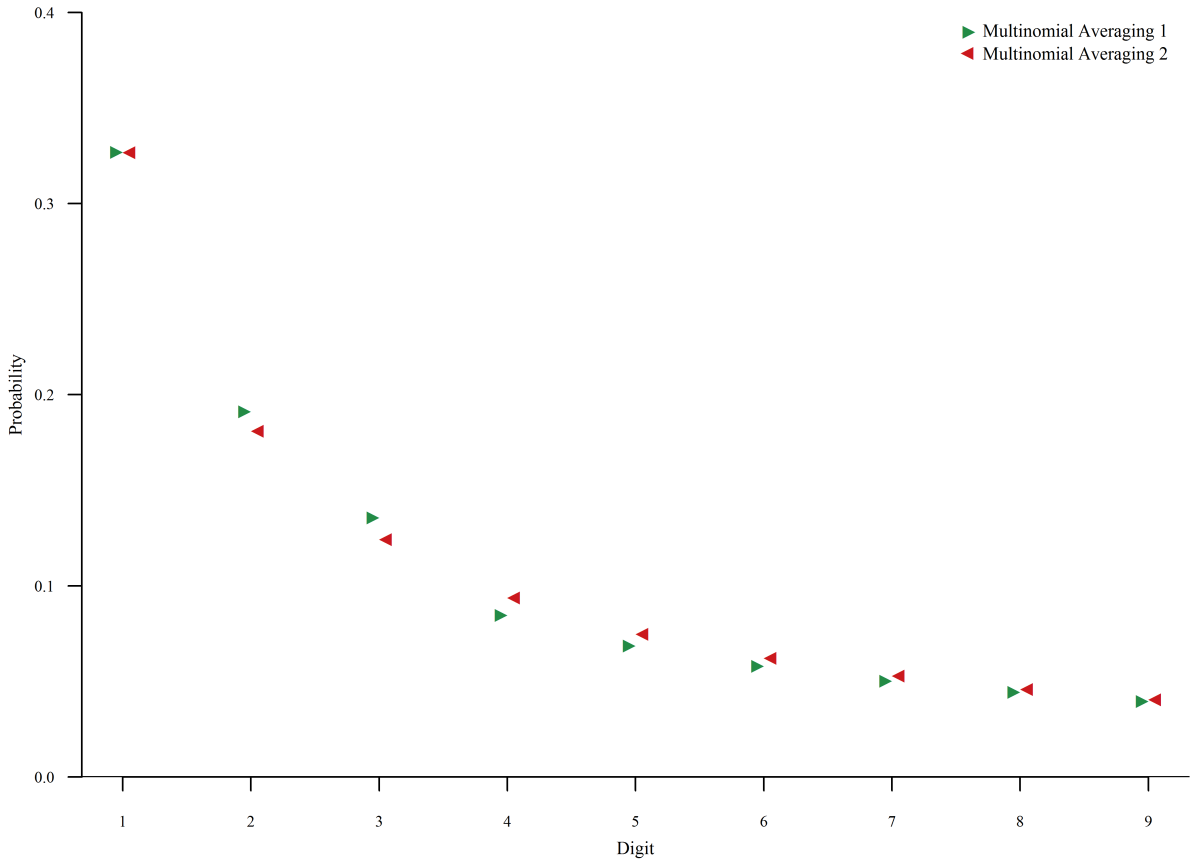


Figure 2.17: *A plot of the expected digit distributions for the two versions of the Multinomial Averaging test, MA1 and MA2. Note that their biggest differences are in digits 2 through 4.*

Figure 2.18 provides the estimated power curves for these two varieties of the Multinomial Averaging test. The MA2 version has higher power. Note, too, that the contamination ranges from 0 to 15%, not to 100%. This test is *also* very sensitive to deviations from the expected digit distribution.

I used both versions of this test on the same 10 elections from earlier (Table 2.5), summarizing the results in Table 2.6. Note that the substantive conclusions are largely the same. The only difference is that the MA1 test no longer finds Norway's election irregular. The two questionable Romanian elections remain highly significant.

Conclusion: In concluding this section, I make the following observations about these three generalized Benford tests. First, the tests are extremely sensitive to violations of

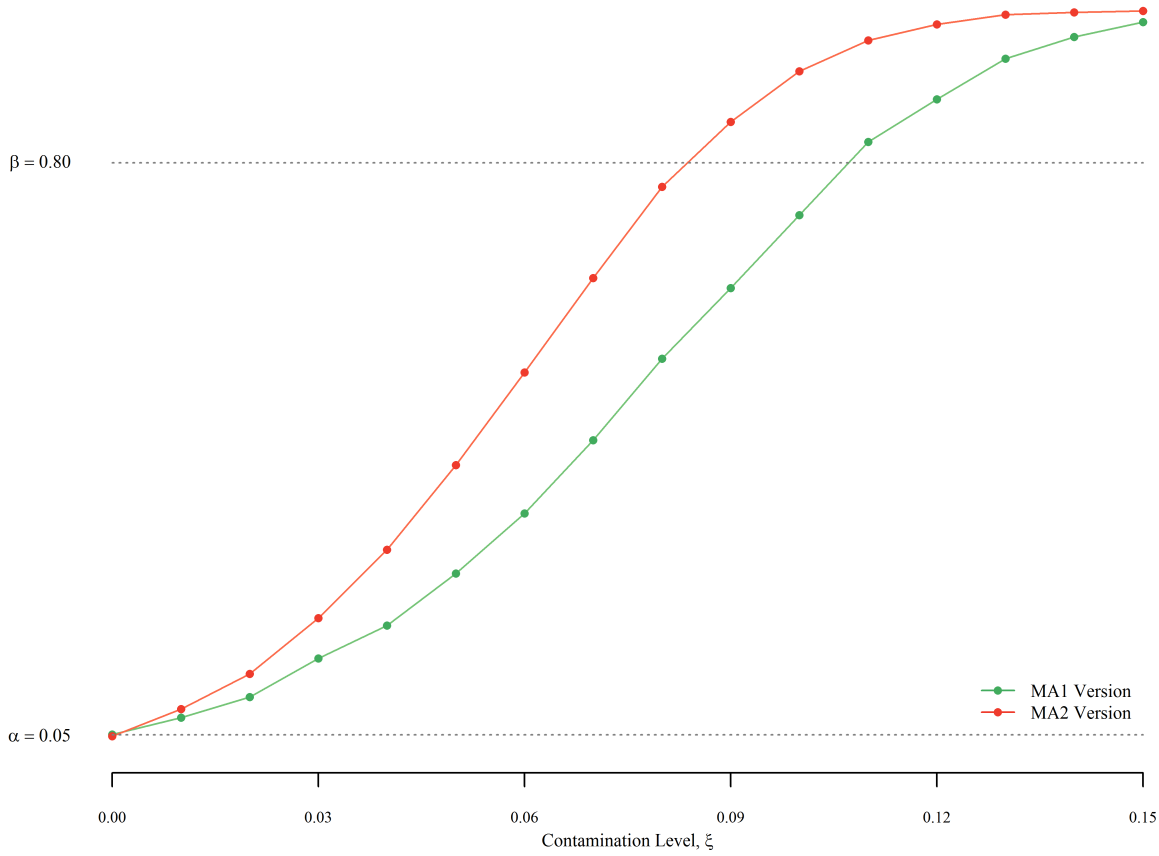


Figure 2.18: *A plot of the power curves for the two versions, MA1 and MA2. Note that the MA2 version has a higher power against shift contamination. The number of iterations at each contamination level is 10,000.*

the assumed null distribution. This means that we are very dependent on knowing the "correct" digit distribution. Assuming that all US states held free and fair elections in 2004 and 2008, the Likelihood Simulation method had a Type I Error rate double what it should have had.

Extending this to several foreign elections, the Likelihood Simulation test rejected three elections that I would contend were also free and fair. Also, while it did flag the two Romanian presidential elections as questionable, the type of fraud alleged in that election is not the type of fraud that this method is designed to detect.

The two Multinomial Averaging tests were also very sensitive to deviations from the expected digit distribution, with MA2 being the more sensitive. The MA1 test flagged the Irish elections, but not the Norwegian general election; the MA2 test flagged all three.

State	Year	Election	MA1	MA2
Afghanistan	2009	President	0.625	0.609
Egypt	2011	Referendum	0.712	0.650
Ireland	2011	Parliament	0.000	0.000
Ireland	2012	Referendum	0.005	0.005
Lithuania	2009	President	0.004	0.003
Macedonia	2011	Parliament	0.559	0.523
Norway	2009	Parliament	0.161	0.035
Romania	2009	President (Round 1)	0.000	0.000
Romania	2009	President (Round 2)	0.000	0.000
United States	2004	President	0.811	0.847

Table 2.6: *Results for the two Multinomial Averaging tests performed on 10 real elections. The MA1 and MA2 columns provide the p-values. Only in the Afghan and Romanian elections was fraud alleged. However, it was not the type that this test is designed to detect.*

Thus, there remain large questions about the applicability of any of these three tests. They are powerful against the null distribution, but questions remain whether that null distribution is the correct distribution. In the next section, we relax the distributional assumption.

2.5.3 The Empirical Benford Tests. In addition to these two generalized Benford tests, I estimate the size and power of the two empirical Benford tests—parametric and non-parametric. These are both Monte Carlo tests, requiring much computing power (and time). Abstractly, the two Monte Carlo tests are the same. They differ only in how the simulated elections are generated. In the parametric version, the election proportions are generated from the Logit-Normal distribution; in the non-parametric version, from the observed proportions.

Parametric Results: To determine the power of this test, I need to first estimate the critical value for the test. The following is the code I used to estimate those critical values.

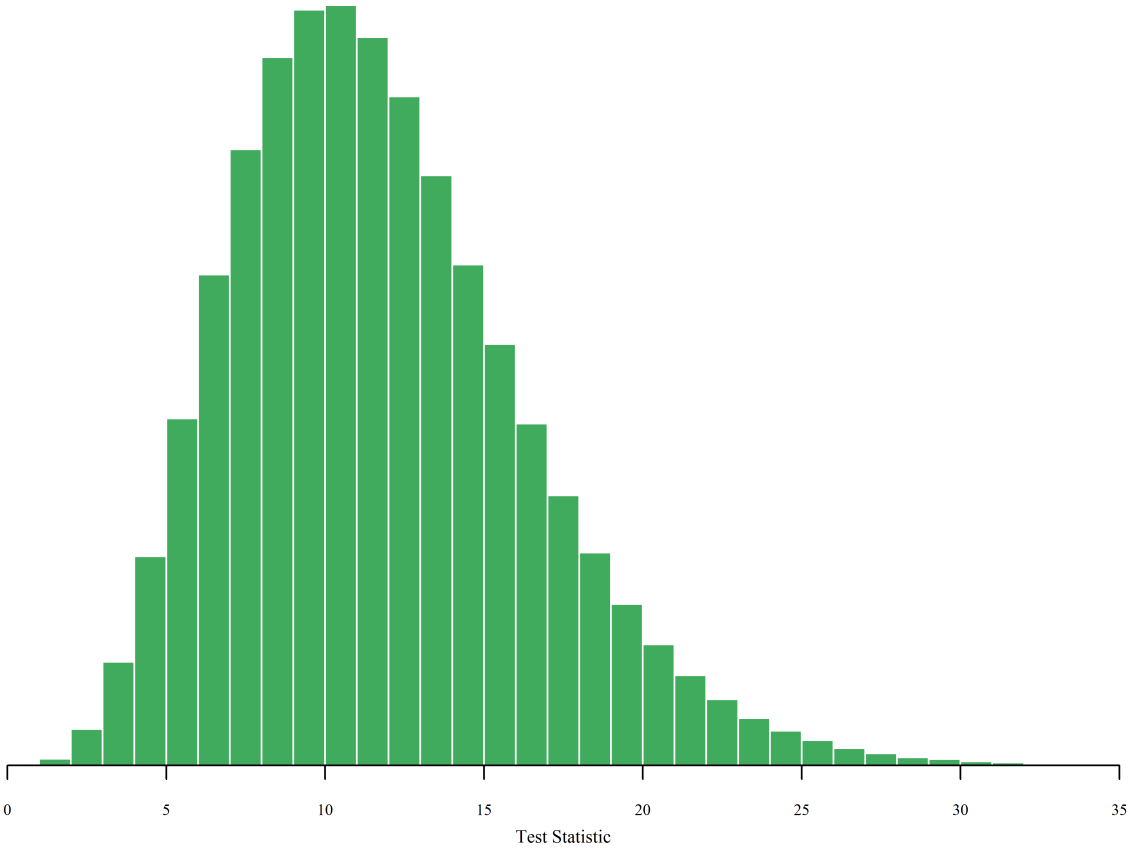


Figure 2.19: *The distribution of the test statistic in the parametric case, based on one million iterations.*

```
usd = read.csv("usa2008pres.csv")
usC = usd$MCCAIN
usN = usd$MCCAIN + usd$OBAMA + usd$OTHER
usP = usC/usN
n   = length(usN)

lgm = mean( logit(usP) )
lgs = sd  ( logit(usP) )

TS = matrix(NA,ncol=1,nrow=1e4)
for(m in 1:1e4) {
  mcP = rlgtnorm(n, m=lgm,s=lgs)
  mcC = mcP*usN
  mcD = lead.digit(mcC)
  mcO = tabulate(mcD,9)
  TS[m] = benford.test(mcO)
}
quantile(TS,c(0.025,0.975))  ## The estimated CVs
```

Figure 2.19 is the estimated distribution of the test statistic in this case. A 95% confidence interval is from 4.36 to 22.30, on one-million iterations. These limits

are close to those estimated in the non-parametric case (below). The main difference between the parametric and the non-parametric script is Line 10. In this case, the election proportions are drawn from a Logit-Normal distribution; in the non-parametric case, from the observed US vote proportions.

There are three variables measured from the US election: candidate vote, total vote, and candidate proportion. The contamination is replacing each leading digit in the candidate vote with a "1." I attempted other types of contamination, but the power of the test for these other contaminations (shift and uniform) was always lower. I varied the contamination level from 0 to 100% in steps of 10%. For each contamination level, I estimated the rejection rate by performing 1000 Monte Carlo tests. Figure 2.20 is a plot of the estimated powers.

Note that the power is never "high." As expected, it takes on a maximum value at 100% contamination, but that value is an anemic 0.17.

Non-Parametric Results: Where the parametric test drew the simulated election proportions from the Logit-Normal distribution, the non-parametric version draws them from the "observed" election. Here, the observed election is the election being tested, which may be contaminated.

Again, the test must be calibrated for the null election, the 2012 US Presidential election. I estimate the critical value with the following code.

```
usd = read.csv("usa2008pres.csv")

usC = usd$MCCAIN
usN = usd$MCCAIN + usd$OBAMA + usd$OTHER
n   = length(usN)
usP = usC/usN
TS = matrix(NA,ncol=1,nrow=B)

for(b in 1:B) {
  elP = sample(usP, n, replace=TRUE)
  elC = floor(elP*usN)
  elD = lead.digit(elC)
  elO = tabulate(elD,9)
  TS[b] = benford.test(elO)
}

quantile(TS,c(0.025,0.975))  ## The estimated CVs
```

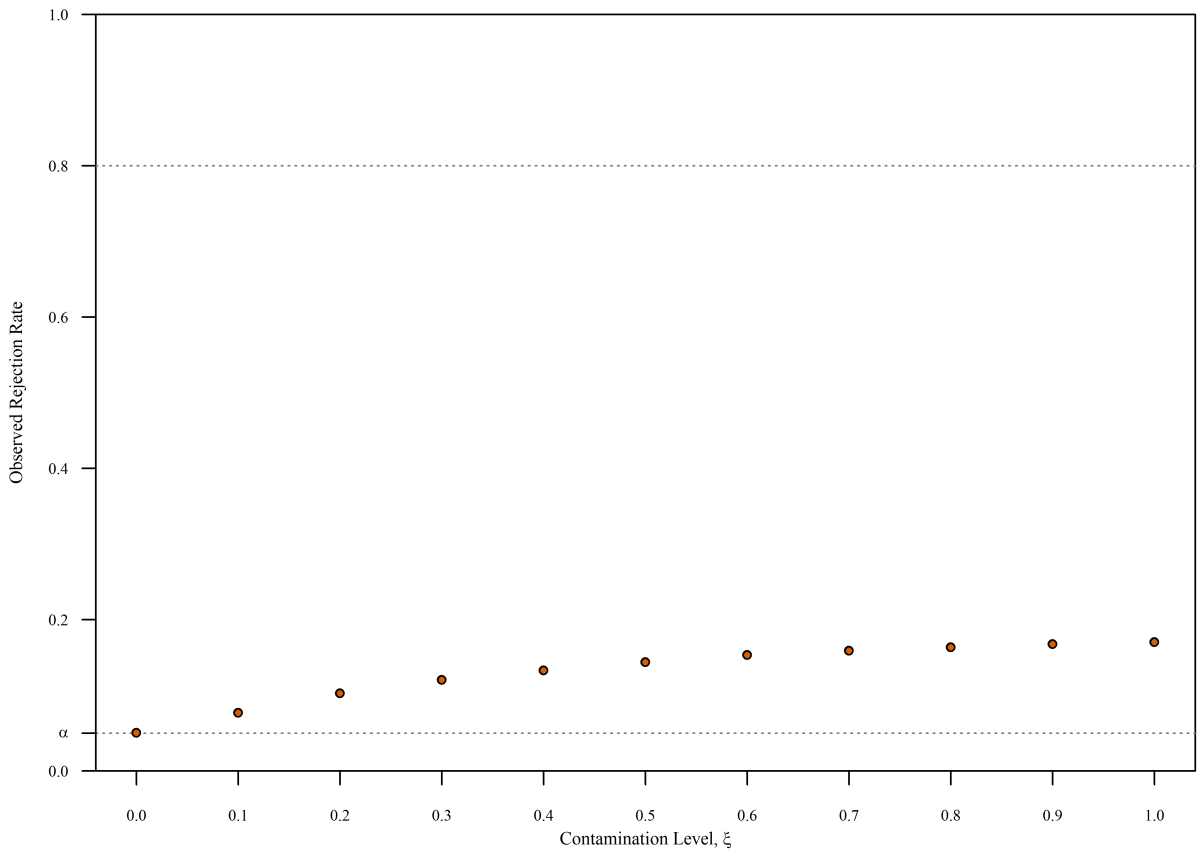


Figure 2.20: *Estimated power curve for parametric empirical Benford test. Note that the rejection rate grows slowly with increasing levels of contamination, ξ. The highest rejection rate is just* 0.17*.*

Figure 2.21 is the estimated distribution of the test statistic in this case. Note that a 95% confidence interval is from 4.44 to 22.23, on one million iterations. These limits are close to those estimated in the parametric case. Again, the main difference between the parametric and the non-parametric case is the simulation line, here Line 8. In the parametric case, the election proportions were drawn from a Logit-Normal distribution; in this case, from the observed US vote proportions.

As in the parametric case, the contamination is of the spike type. Full contamination has all leading digits a "1." As above, various levels of contamination are achieved by mixing the fully-contaminated and the null distributions.

Figure 2.22 is the estimated power curve for the non-parametric version of the empirical Benford test. As in the parametric case, the power curve is rather flat and

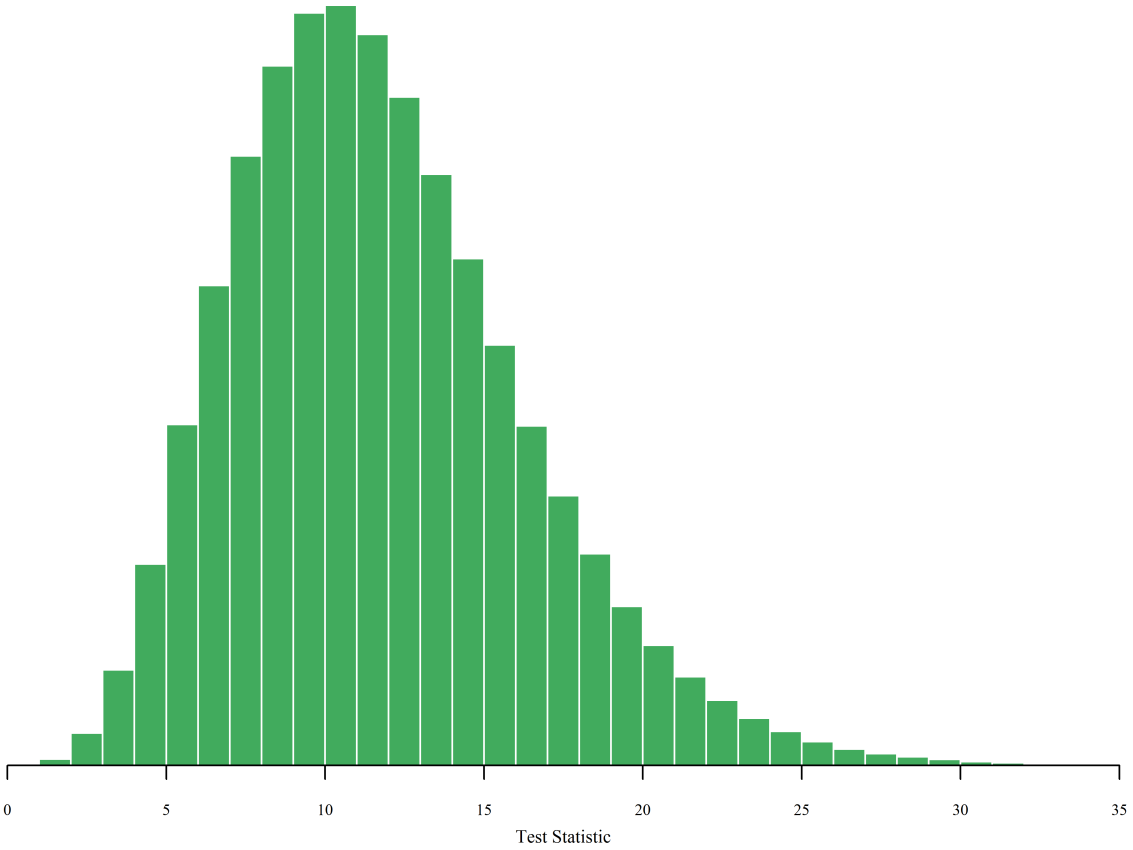


Figure 2.21: *The distribution of the test statistic in the non-parametric case, based on one million iterations.*

disappointing. At full contamination, the rejection rate is only 0.189, meaning that this procedure is only able to detect a fully-contaminated election 19% of the time.

As in the parametric case, other contamination schemes did not fare as well. Both uniform contamination and shift contamination had power curves below this.

Again, we are faced with a disappointing conclusion. From the above analysis, I cannot fully recommend any of these tests. While the Type I Error rates are close to nominal, owing to the critical values being set to ensure that happened, the tests' powers are rather feeble.

This raises the problem of why an election fails the test, were it to fail. Would it fail because it was unfair, or would it fail because of the natural Type I Error rate? When power is this low, it is difficult to distinguish a false positive from a true positive.

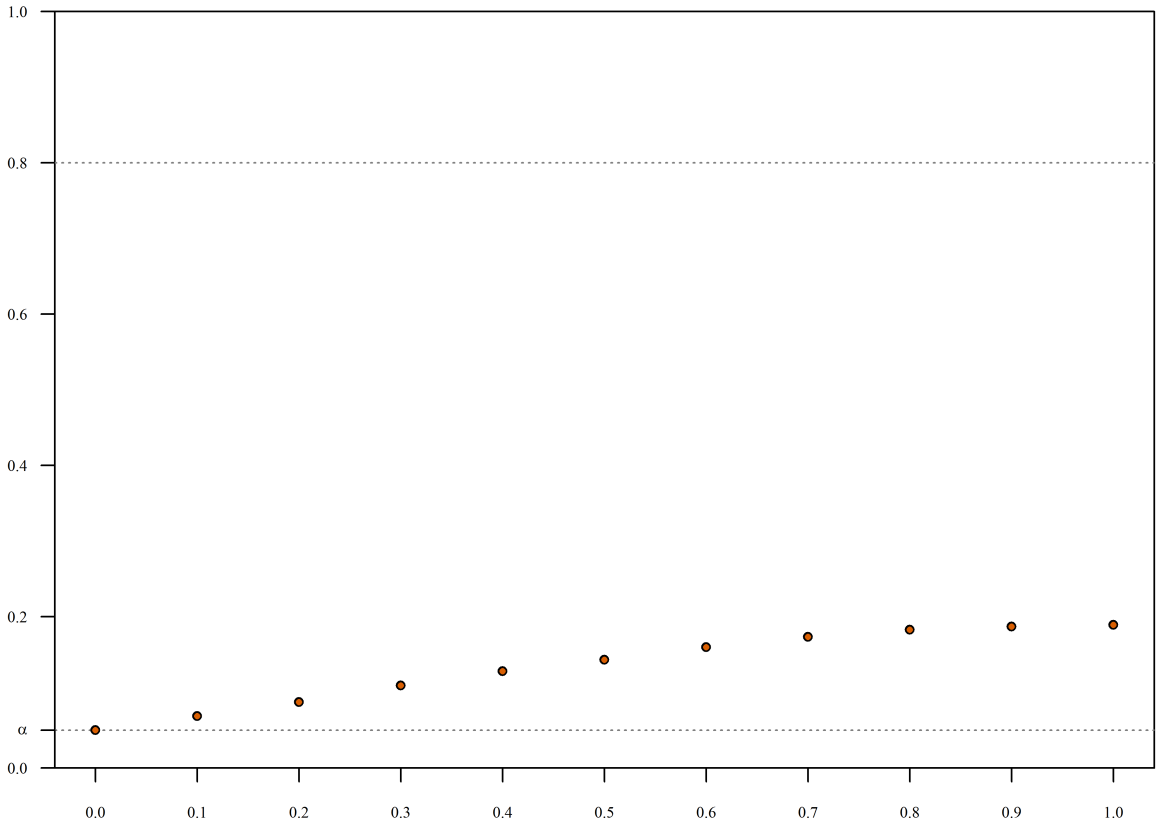


Figure 2.22: *Estimated power curve for non-parametric empirical Benford test. Note that the estimated rejection rate remains low throughout. The maximum power is reached at full contamination, with a rate of just* 0.19*.*

2.6. Conclusion

Frequently, governments release vote counts at the division level. The counts have a distribution. Following the observation that the first several pages in a book of logarithm tables were more used that later pages, both Newcomb (1881) and Benford (1938) showed that the leading digit of certain types of data had a specific distribution—eponymously called the Benford distribution. Its use in electoral forensics, however, makes the assumption that the division sizes are exact powers of 10, which is not the case (Mebane 2010). To fix this, I introduced and explored what I termed the generalized Benford distribution.

While the distribution may work well for testing the leading digits of counts from electoral divisions of equal size, it is not directly applicable for electoral divisions of different sizes. The rest of the chapter explored several tests that could be used to handle such cases, as well as tests to determine if this distribution is applicable to elections in general.

The two generalized Benford tests I examined estimated the final digit distribution either using simulation or using an averaging method. The former is the Likelihood Simulation test; the latter, the two Multinomial Averaging methods. In addition to the generalized Benford tests, I formulated two empirical Benford tests. The first is a parametric method that assumes the vote proportion follows a Logit-Normal distribution. The second is a non-parametric method that estimates the p-value and confidence intervals using a bootstrap method.

In the cases of the generalized Benford test, the Likelihood Simulation method produced inflated Type I Error rates—double expected. The two Multinomial Averaging methods produced results quite similar to those of the Likelihood Simulation method. In all three cases, the power curve showed a very high sensitivity to deviations from the expected digit distribution. As we are not truly sure about the correct distribution of count digits, this sensitivity may have been the cause of the inflated Type I Error rate.

The two empirical Benford tests showed a different problem. While the empirical Type I Error rates were approximately nominal, their power curves were relatively flat with estimated maximum power of 0.17 (parametric test) and 0.19 (non-parametric test). Both are rather low.

Because of these results, I am able to recommend the use of the generalized Benford test *as long as* it is just one of several tests used. The three tests were very sensitive to departures from the assumed distribution, thus these tests may be treated as "gatekeeper" tests—if the election passes, there is no real evidence of this type of unfairness (vote count fabrication). If the election fails, further tests are needed.

While these results are rather disappointing, they are not entirely surprising. Recall from Section 2.3.1 that the Benford distribution is equivalent to making the assumption the vote proportions are Uniformly distributed. As this assumption is not met in reality, one would not expect the Benford test *in any of its guises* to be appropriate.

In the next section, we move on from digit tests to regression tests. In this chapter, we only had vote counts to analyze. In the next chapter, we include invalidation rates. Under the free and fair hypothesis, these two variables should be independent. Thus, Chapter 3 examines how to best test this assumptions, even in the presence of electoral fraud.

2.7. Annex 1

The R code for the Westfall and Wolfinger (1997) min-p test.

```
minp.test = function(obs,exp, table=TRUE) {
  # Call table function
  source("minp.table.R")
  tbl = minp.table(obs,exp)

  # Initialize variables
  n    = sum(obs)
  pval = numeric()
  pit  = numeric()

  # Get pj = min.p
  pj=1.00
  for(d in 1:9) {
    pp=tbl[obs[d]+1,d]
    if( pp<pj ) pj=pp
  }

  # Calculate p_{it(j)}
  pit = numeric()
  for(d in 1:9) {
    st = sort( as.numeric( tbl[,d]) )
    wo = which( st<=pj )
    if( length(wo)>0 ) {
      to = max( wo )
      pit[d] = st[to]
    } else {
      pit[d]=0
    }
  }

  # Calculate the adjusted p-value
  padj = 1 - prod(1-pit)

  names(obs)=1:9
  names(exp)=1:9
  res = list(source="Westfall and Wolfinger 1997")
  if(table) {
    res$prop.table=tbl
  }
  res$observed = obs
  res$expected.freq = exp
  res$expected.vals = exp*n
  res$pit = pit
  res$min.p = pj
  res$bonferroni.adj.p.val = min(1,9*pj)
  res$ww.adj.p.val = min(padj,1)

return( res )
}
```

The R code for the `minp.table` function.

```
minp.table = function(obs,exp) {
 tmax = sum(obs)
 ptop = numeric()
 pbot = numeric()

 pvals = matrix(NA, nrow=tmax+1,ncol=9 )
 dimnames(pvals) = list(freq=0:tmax,digit=1:9)

 mi = tmax * exp

 for(d in 1:9) {
  for(t in 0:tmax) {
   r = 2*mi[d] - t

   ptop =   pbinom(t,  size=tmax,prob=exp[d] ) +
          1-pbinom(r,  size=tmax,prob=exp[d] )
   if(is.integer(r)) ptop=ptop+dbinom(r,size=tmax,prob=exp[d])

   pbot =   pbinom(r,  size=tmax,prob=exp[d] ) +
          1-pbinom(t,  size=tmax,prob=exp[d] ) +
            dbinom(t,  size=tmax,prob=exp[d] )

  pvals[t+1,d] = min(ptop,pbot)
  }
 }
return( round(pvals,3) )
}
```

2.8. Annex 2

The R code for creating B simulated elections:

```
## Import the election
us2008 = read.csv("usa2008pres.csv")

# Extract the needed data
divSize = us2008$OBAMA + us2008$MCCAIN + us2008$OTHER
divProp = us2008$MCCAIN/divSize

# Initialize the variables
B = 10000
elections = matrix(NA, nrow=B, ncol=length(divProp))

# Begin: loop
for(i in 1:B) {
  perm = sample(length(divProp))
  elections[i,] = floor(divProp[perm]*divSize)
}
# End: loop
```

CHAPTER 3

REGRESSION TESTS

Côte d'Ivoire, 2010. After almost a decade of civil war and five years of postponed elections, Ivoirians went to the polls in 2010 to elect their fifth president. Three major candidates presented themselves for the first round: Henri Bédié, the republic's second president who ascended to the position after the death of long-time leader Félix Houphouët-Boigny; Alassane Ouattara, a former prime minister who left his position in 1993 at the urging of Bédié; and President Laurent Gbagbo, whose presidency spanned the civil war that embroiled the country for eight years.

October 31 was the date for the first round. Should no candidate obtain 50%+1 of the votes cast, there would be a runoff election between the two leading vote recipients. As expected, militias and the remnants of the northern rebellion made voting in the west and the north dangerous. Candidates accused each other of vote fraud. Multiple sources reported different vote totals for these three candidates. The official source, the Independent Electoral Commission (CEI), gave Bédié third place with only 25.2% of the vote (Yàn 2010).

The November 28 second round election featured President Gbagbo and Ouattara. Ouattara accused Gbagbo of causing the civil war, while Gbagbo claimed Ouattara had planned two coup attempts. Supporters of each were divided along tribal lines.

The CEI began to announce tallies as they became available. However, the country's division found its way even to the allegedly independent CEI, with one member snatching official vote counts from another during a press conference. No results were announced that night.

Each side accused the other of vote fraud. Security forces sought to stifle the violence, but merely added to it. By the time the CEI announced that Ouattara won the election with 54% of the vote, the two sides were in open conflict. To make matters worse, the Constitutional Council declared the CEI had no legitimacy and announced President Gbagbo had won the election (BBC News 2011).

Although the international community recognized Ouattara as the legitimate president, it took foreign intervention and six months of civil war before Gbagbo left the presidency, allowing Ouattara to take the helm of the deeply divided country (BBC News 2011).

And yet, the question remains. Two official Ivoirian agencies reported different election outcomes. The Constitutional Council claimed the CEI returns were fraudulent. To rectify this, the Constitutional Council invalidated all votes in the seven northern provinces—Ouattara's stronghold (AFP 2010).

Did the Constitutional Court have evidence of electoral fraud? If so, what was it?

3.1. Introduction

In the previous chapter, I explored several tests of the reported vote count. Since that was the only information available, those tests compared the observed count to a hypothesized distribution. In that chapter, we discovered that the tests were of questionable utility. Either the hypothesized distribution did not appear to match reality or the test was of low power.

In this chapter, I examine tests comparing the invalidation rate with the candidate support rate. Under the free and fair hypothesis, these two variables should be independent. If not, then the invalidation rate depends on for whom the ballot was cast. A violation of the free and fair hypothesis may be due to an unfairness in the electoral system or to electoral fraud.

An example of the former: the elderly support Candidate X over Candidate Y, but they tend to fill out their ballots improperly. This increases the invalidation rate for Candidate X. The tests of this chapter will detect that unfairness if the elderly tend to live in certain electoral divisions over others.

An example of the latter: Supporters of Candidate X stuff the ballot box in certain electoral divisions. Those ballots have two things in common: They are filled out correctly, and they are filled out for Candidate X. This can be detected using the methods of this chapter as the two variables, invalidation rate and candidate support rate, are no longer independent.

These tests are all variations of regression tests. As such, I begin with a review of regression schemes, examining ordinary least squares (OLS) regression and its assumptions, which are not met in this type of environment. Weighted least squares (WLS) takes care of some of the problems with OLS, but WLS assumes the covariance matrix is known. Here, that is not the case, and I propose feasible generalized least squares regression (FGLS) as the solution to estimating the effects of the candidate support on the invalidation rate.

Secondly, I note that there may be two types of electoral divisions: fair and unfair. Current methods in electoral forensics treat all divisions as though they were from the same population. I hypothesize a threshold τ that separates the fair from the unfair. The key is to estimate it correctly. To solve this problem, I introduce three options: threshold search, the Expectation-Maximization algorithm (as the Healy-Westmacott algorithm), and an empirical version of Bayesian regression.

3.2. Least Squares Regression

The previous chapter dealt with the case in which the government only reports the vote counts for the candidates at the electoral division level. There, I used digit tests to examine the likelihood that the data were generated from a natural voting process.

In addition to the candidate vote counts, governments may also report the number of invalidated ballots at the electoral division level. In such cases, we can perform additional tests.

In the presence of invalidation, the free and fair hypothesis is that each person's vote has the same probability of being invalidated as any other person's ballot—this invalidation must be independent of the candidate chosen on the ballot.

3.2.1 Ordinary Least Squares. A typical test for independence of two ratio-level variables is a t-test on the estimated slope parameter arising from ordinary least squares regression (OLS). The two most important strengths of using OLS to estimate the effect are that OLS is robust to violations of its assumptions and that OLS is easily performed.

However, violations of the homoskedasticity assumption produce biased standard error estimates, which means the p-values are biased. Solutions to this issue include transforming the variables (Younger 1979) and adjusting the estimated standard errors (White 1980). In lieu of adjusting the data or the standard error estimates, one can use knowledge about the process to improve the model.

As the dependent variable is a proportion, the expected variance of the distribution of the observations under the null hypothesis is $\pi(1-\pi)/N_i$, where N_i is the number of votes cast in electoral division i. Note that the variances are equal under the null hypothesis if *and only if* the division sizes are equal. As division sizes tend to be unequal, the model may be heteroskedastic. A usual solution to this issue is to use weighted least squares to estimate the slopes.

In matrix form, the linear model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Y}$ is the column vector of invalidation rates, $\mathbf{X}$ is the design matrix consisting of a column of 1's and a column of candidate support rates, $\boldsymbol{\beta}$ is the vector of effect sizes, and $\boldsymbol{\varepsilon}$ is the vector of residuals. Under ordinary least squares, we assume $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$, which implies

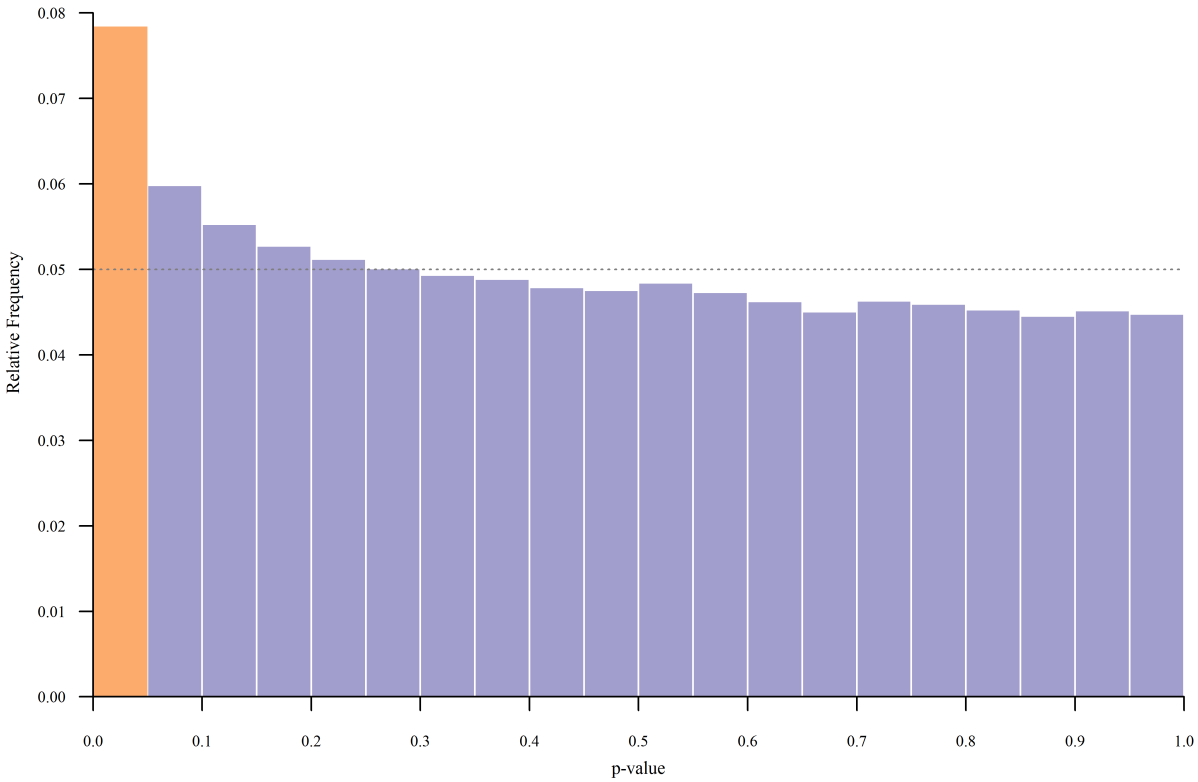


Figure 3.1: *Histogram of p-values for using OLS under the null hypothesis. The observed rejection rate (height) at the* $\alpha = 0.05$ *level is* 0.07837 *instead of* 0.05*. Note that this observed rejection rate is based on 100,000 trials.*

$\mathbf{Y} \,|\, \mathbf{X} \sim \mathcal{N}_n \left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n \right)$. The OLS estimates are

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Y} \tag{3.1}$$

From this, we know

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} \sim \mathcal{N} \left(\boldsymbol{\beta}, \left(\mathbf{X}'\mathbf{X} \right)^{-1} \sigma^2 \right),$$

which produces the expected hypothesis tests and confidence intervals.

3.2.2 Weighted Least Squares. While Equation 3.1 produces unbiased estimates of $\boldsymbol{\beta}$ in the absence of misspecification, those estimates are inefficient in the presence of heteroskedasticity (Kennedy 2003). This results in improper rejection rates under the null hypothesis. Figure 3.1 is a histogram of the p-values under the null hypothesis, with

the height of the first bar representing the rejection rate at $\alpha = 0.05$. For the p-value to have its usual meaning, its distribution must be Uniform(0,1). For this histogram, that is not the case. In fact, the usual OLS test rejects at a 0.07837 rate.

In terms of the current paradigm, the model remains $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is an arbitrary, yet known, positive definite covariance matrix. It is trivial to show (e.g., Christensen 2002) that, under the assumption that the errors are Normally distributed

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = \left(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$
$$\hat{\boldsymbol{\beta}}_{\text{WLS}} \sim \mathcal{N}\left(\boldsymbol{\beta},\ \left(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\right)$$

In this research, the null hypothesis is that the errors are independent; that is, $\boldsymbol{\Sigma}$ is diagonal. Specifically, $\boldsymbol{\Sigma} = \sigma^2\pi(1-\pi)diag\left\{\frac{1}{N_i}\right\}$. Thus, the distribution of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} \sim \mathcal{N}\left(\boldsymbol{\beta},\ \sigma^2\pi(1-\pi)\left(\mathbf{X}'diag\left\{N_i\right\}\mathbf{X}\right)^{-1}\right),$$

where σ^2 is a measure of overdispersion.

Since we do not know the value of π, it must be estimated from the data. Unfortunately, the weighted least squares method will produce biased standard error estimates in this case (Fromby et al. 1984). Note that this bias is most severe when the variance is a function of an predictor variable. To produce unbiased estimates of the standard errors when the covariance matrix is unknown, one can use feasible generalized least squares (FGLS).

3.2.3 Feasible Generalized Least Squares. Because the weighted least squares method requires the covariance matrix be known, it is unsuitable for examining elections in this context. The covariance matrix depends on the estimated parameter π. To fix this issue, Fromby et al. (1984) created the feasible generalized least squares method.

This method is frequently defined in the literature (see, e.g., Chaturvedi 1995; Fromby et al. 1984; Klaassen and Magnus 2001; Magee 1998). The methods appear to

agree on one thing: FGLS estimates the covariance matrix using ordinary least squares regression, then uses that matrix in weighted least squares regression.

Following the work of Fromby et al. (1984), the FGLS routine I use consists of iterating two steps until convergence: fit the data with the current estimated covariance matrix, estimate the covariance matrix using the residuals (Gullickson 2007).

Under the null hypothesis, the electoral problem provides structure to the covariance matrix. The invalidation rates are independent across the electoral divisions. The population invalidation rate is constant across the electoral divisions. Under these assumptions, the covariance matrix will be

$$\mathbf{\Sigma} = \begin{bmatrix} \frac{\sigma^2\pi(1-\pi)}{N_1} & 0 & 0 & \cdots & 0 \\ 0 & \frac{\sigma^2\pi(1-\pi)}{N_2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{\sigma^2\pi(1-\pi)}{N_3} & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{\sigma^2\pi(1-\pi)}{N_n} \end{bmatrix} = \sigma^2\pi(1-\pi)\mathbf{N}^{-1}$$

Both σ^2 and π need to be estimated from the data. In this context, σ^2 is an overdispersion parameter, π is the population invalidation rate, and $\mathbf{N} = diag\{N_1, N_2, \ldots, N_n\}$, with N_i the vote total in electoral division i.

In the OLS step, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, and $\hat{\beta}_0$ serves as the current estimate of π. The current estimate of σ^2 comes from the usual formula for the observed covariance matrix

$$\hat{\sigma}^2 = \frac{1}{(n-2)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\frac{N_i}{\hat{\pi}(1-\hat{\pi})}$$

This formula arises from the usual estimator of the diagonal elements in the covariance matrix

$$\hat{\mathbf{\Sigma}}_{ii} = \frac{1}{(n-2)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

solved for $\hat{\sigma}^2$, where $\hat{\mathbf{\Sigma}}_{ii} = \hat{\sigma}^2\ \frac{\hat{\pi}(1-\hat{\pi})}{N_i}$.

In subsequent steps, the estimate uses the previous estimate of the covariance matrix. Thus, the kth iteration is

$$\begin{aligned}
&\text{Step 1: } \hat{\boldsymbol{\beta}}^{(k)} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1(k-1)}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1(k-1)}\mathbf{Y} \\
&\text{Step 2: } \pi^{(k)} = \hat{\beta}_0^{(k)} \\
&\text{Step 3: } \hat{\sigma}^{2(k)} = \frac{1}{n-2}\,\frac{(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{N}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{\pi^{(k)}(1-\pi^{(k)})} \\
&\text{Step 4: } \boldsymbol{\Sigma}^{(k)} = \hat{\sigma}^{2(k)}\pi^{(k)}(1-\pi^{(k)})\mathbf{N}
\end{aligned}$$

Iteration continues until the change in $\hat{\sigma}^{2(k)}$ is sufficiently small.

Using this procedure, Fromby et al. (1984) showed the parameter estimates have the following asymptotic distribution $\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \hat{\boldsymbol{\Sigma}}\right)$, with

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Y}, \text{ and } \hat{\boldsymbol{\Sigma}} = \hat{\sigma}^2\hat{\pi}(1-\hat{\pi})\mathbf{N}$$

Here, the hats represent the values of the estimates at the final iteration.

Compare Figure 3.2 to Figure 3.1, which both consist of 100,000 p-values calculated from data generated under the null hypothesis and the two fitting schemes. Under ordinary least squares (Figure 3.1), the actual rejection rate was more than 50% higher than it should have been. Under feasible generalized least squares, the observed rejection rate is within the 95% confidence interval of $\alpha = 0.05$. This indicates FGLS is a better routine than OLS, at least under this metric.

3.2.4 An Illustration. To illustrate feasible generalized least squares regression with election data, let us return to the Afghan presidential election of 2009. Table 3.1 provides both the OLS and the FGLS estimates of the model predicting the invalidation rate using the proportion of the vote for Hamid Karzai.

The OLS symmetric 95% confidence interval for the candidate effect β_1 is from -0.0548 to 0.0523, with a p-value for the non-directional hypothesis of 0.9611. The FGLS symmetric 95% confidence interval is from -0.0004 to 0.0145, with a p-value for

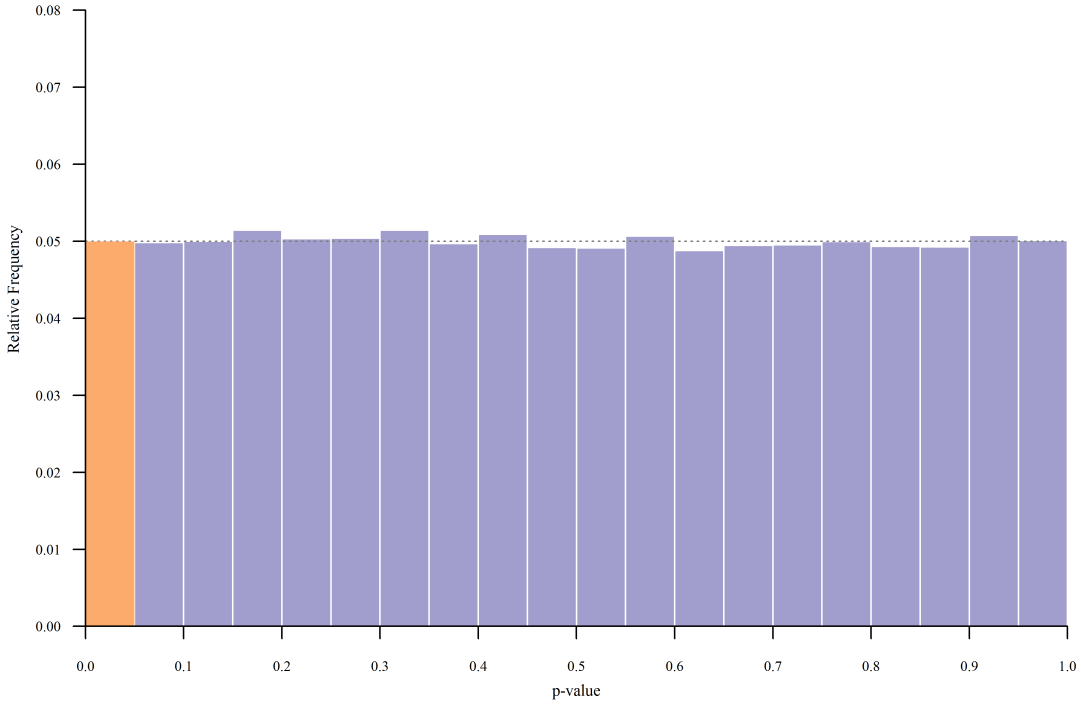


Figure 3.2: *Histogram of p-values for using FGLS under the null hypothesis. Note that the observed rejection rate (orange bar) is* 0.04994*, which is well within a 95% confidence interval of* $\alpha = 0.05$ *(trials=100,000). Thus, FGLS appears to fix problems with OLS.*

the non-directional hypothesis of 0.0628. While we are still unable to reject the free and fair hypothesis at the usual level, it is much closer. This difference is due to the covariance matrix $\hat{\mathbf{\Sigma}}$ estimated in Step 4.

Three final notes before continuing. First, FGLS is an iterative routine, while OLS and WLS are not. This makes FGLS slower than either of the other two methods. In fact, FGLS is OLS plus iterations of WLS. Second, because the variance is not a function of an independent variable, there is little difference in the parameter estimates *or their standard errors* produced by WLS and by FGLS. Third, because of these two observations, I suggest using weighted least squares, unless the calculated p-values are close to α.

Parameter	OLS	WLS	FGLS
π_0	0.0500	0.0542	0.0554
β_1	-0.0013	-0.0060	-0.0061
σ^2	0.0266	1308.32	1160.75

Table 3.1: *Comparison of parameter estimates for the 2009 Afghan presidential election. Note that the parameter estimates using WLS and FGLS are quite similar.*

3.3. Changepoint Regression

Each of the previous regression methods assumed the data came from a single population. That is, they assumed the data-generating process was the same across *all* electoral divisions.

However, in the presence of violations of the free-and-fair hypothesis, this may not be the case. In those countries that do not count the votes centrally, electoral fraud may take place differentially across the electoral divisions; some divisions experience it, others not.

Figure 3.3 illustrates the problem that arises because of the presence of two populations. The points represent each electoral division. There are two types: those from electoral divisions experiencing electoral fraud (orange) and those not (green). If all electoral divisions are treated as originating from a single population (left panel), the effect is not statistically significant ($p = 0.121$). If the two populations are treated separately, the regression on the orange sample *does* detect a statistically significant relationship ($p = 0.029$).

The regression method used to regress on two separate populations goes by many names: hockey-stick, piecewise, broken-stick, etc. (Yanagimoto and Yamamoto 1979). This regression is a type of non-linear regression.

For this problem, the statistical model is

$$\mathbf{Y} = \alpha_0 \mathbb{1}\{\mathbf{X} < \tau\} + \beta_0 \mathbb{1}\{\mathbf{X} \geq \tau\} + \beta_1 \mathbf{X} \mathbb{1}\{\mathbf{X} \geq \tau\} + \boldsymbol{\varepsilon} \tag{3.2}$$

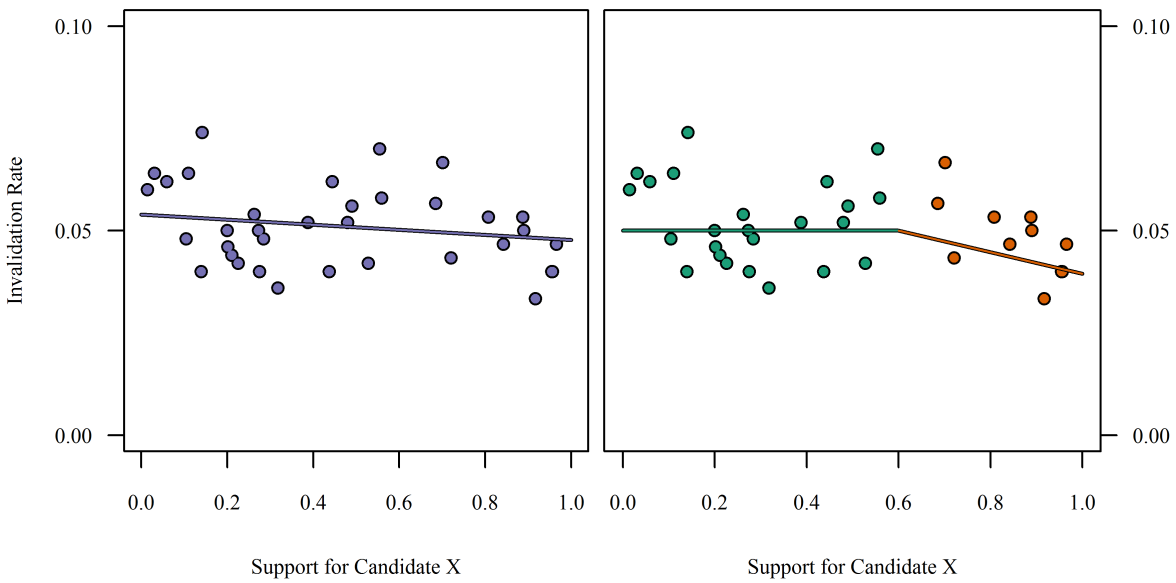


Figure 3.3: *Graphics demonstrating the need to be able to test for two populations. In the left panel, the single regression line on the left does not have a statistically significant slope. The right panel, the slope of the post-threshold regression line is statistically significant at the* $\alpha = 0.05$ *level.*

Here, τ is the cutpoint (threshold) separating the two populations. This model explicitly assumes that all electoral divisions with candidate support less than τ are fair. With this model, the null hypothesis is $H_0 : \beta_1 = 0$ and $\alpha_0 = \beta_0$. If τ is known in advance, this model can be fit with any regression method from the previous section. However, if τ must be estimated from the data, the solution is not as easy.

In the remainder of this section, I explore three methods for detecting the cutpoint of the two populations: threshold grid search, Healy-Westmacott regression, and Bayesian regression.

3.3.1 Threshold Search. The usual situation is that the threshold τ is not known *a priori*; it must be estimated from the data. One method is to estimate τ from the data by selecting multiple thresholds, measuring a relevant quantity at each, and selecting the threshold based on those measurements. As regression often seeks to minimize the sum of the square residuals, this is an appropriate relevant quantity.

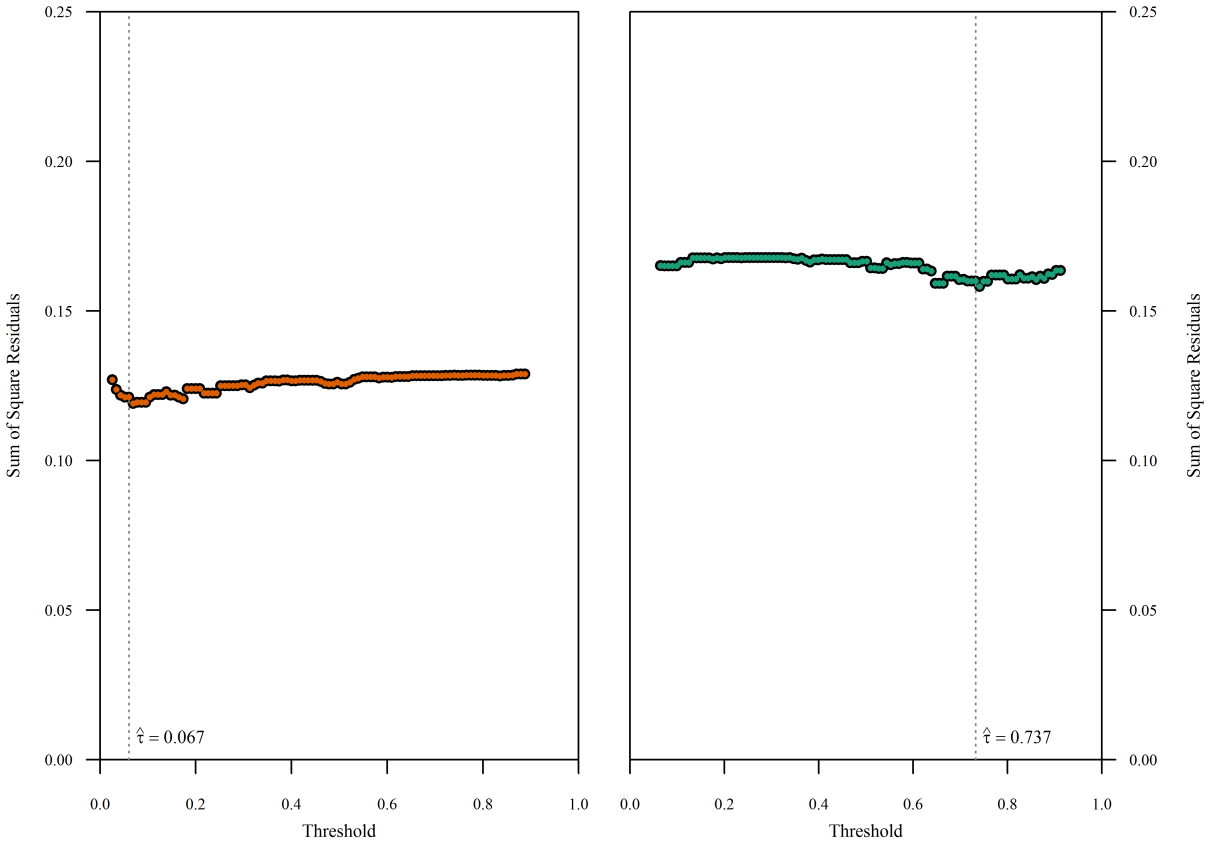


Figure 3.4: *Plots of the sum of square residuals against the threshold, τ. The left panel is the plot for the results of the 2010 Ivoirian runoff presidential election. The right panel is the plot for independent data under similar constraints. Note that there appears to be little difference between the two plots.*

Figure 3.4 (left panel) provides a scatter plot of the sum of square errors against the threshold τ for the 2010 Ivoirian presidential runoff election. Note that each possible threshold produces approximately the same error level. In other words, there is no strong evidence for one threshold over another.

This is echoed in the independent case (Figure 3.4, right panel). The elections for this case were generated to have the two variables, invalidation rate and candidate support rate, independent of each other, but to match the Ivoirian election in all other ways. Note that it, too, is relatively flat. These two graphics strongly suggest that there is no boundary between fair divisions and unfair divisions. All electoral divisions are equally fair (or unfair).

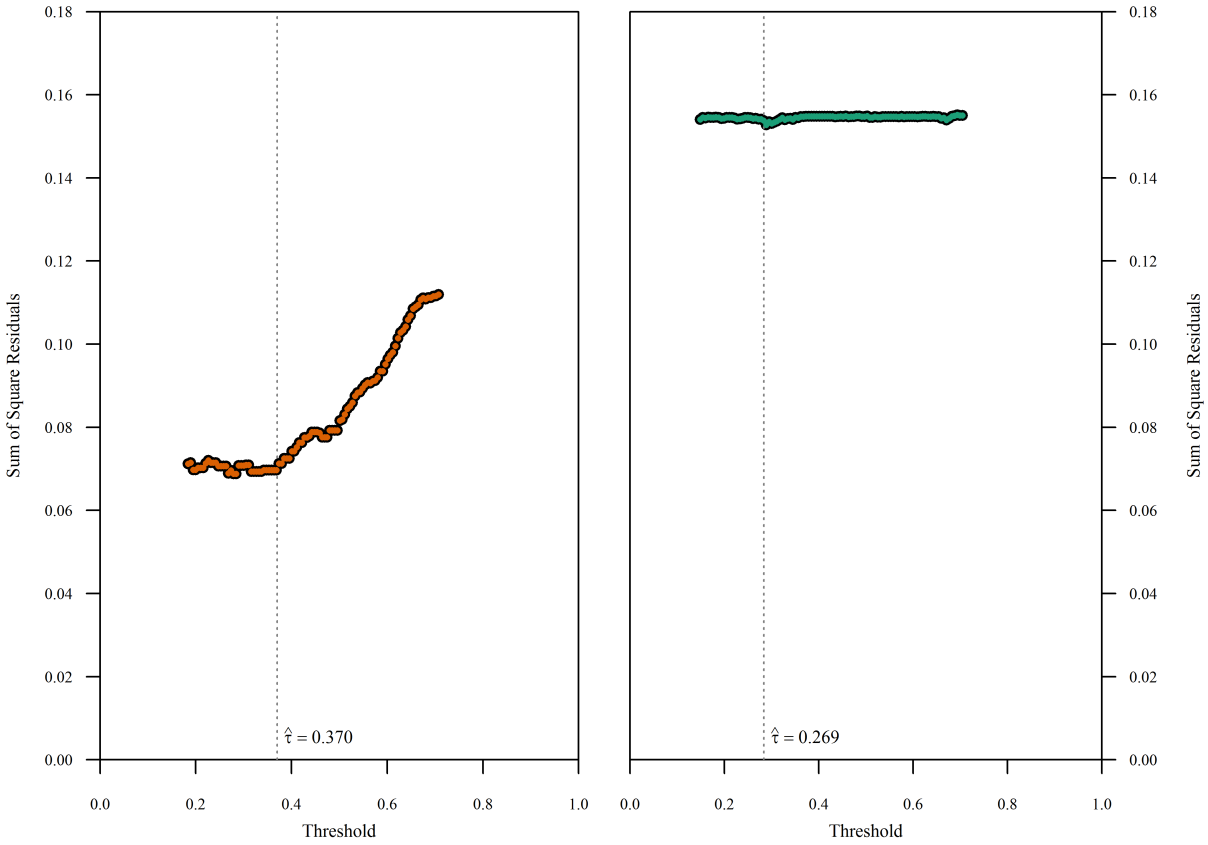


Figure 3.5: *Plots of the sum of square residuals against the threshold, τ. The left panel is the plot for the results of the 2010 Sri Lankan presidential election. The right panel is the plot for independent data under similar constraints. Note the difference in form between the two plots.*

Contrast this with Figure 3.5 (left panel), which provides a plot of the sum of square errors against threshold value for the 2010 Sri Lankan presidential runoff election. Note that those thresholds left of $\tau = 0.37$ do not affect the mean square error of the model much; however, those thresholds to the right of τ do. This suggests that those electoral divisions which had a vote of less than 0.37 for President Mahinda Rajapaksa had a relatively constant invalidation rate, while those with a higher vote proportion for the president did not. This is evidence against the free and fair hypothesis.

Figure 3.5 (right panel) provides a plot of the sum of square errors against threshold value for the independent case. These data were created from two independent uniform distributions. The x-values ranged between the maximum and the minimum vote share for Rajapaksa. The y-values ranged between the maximum and minimum invali-

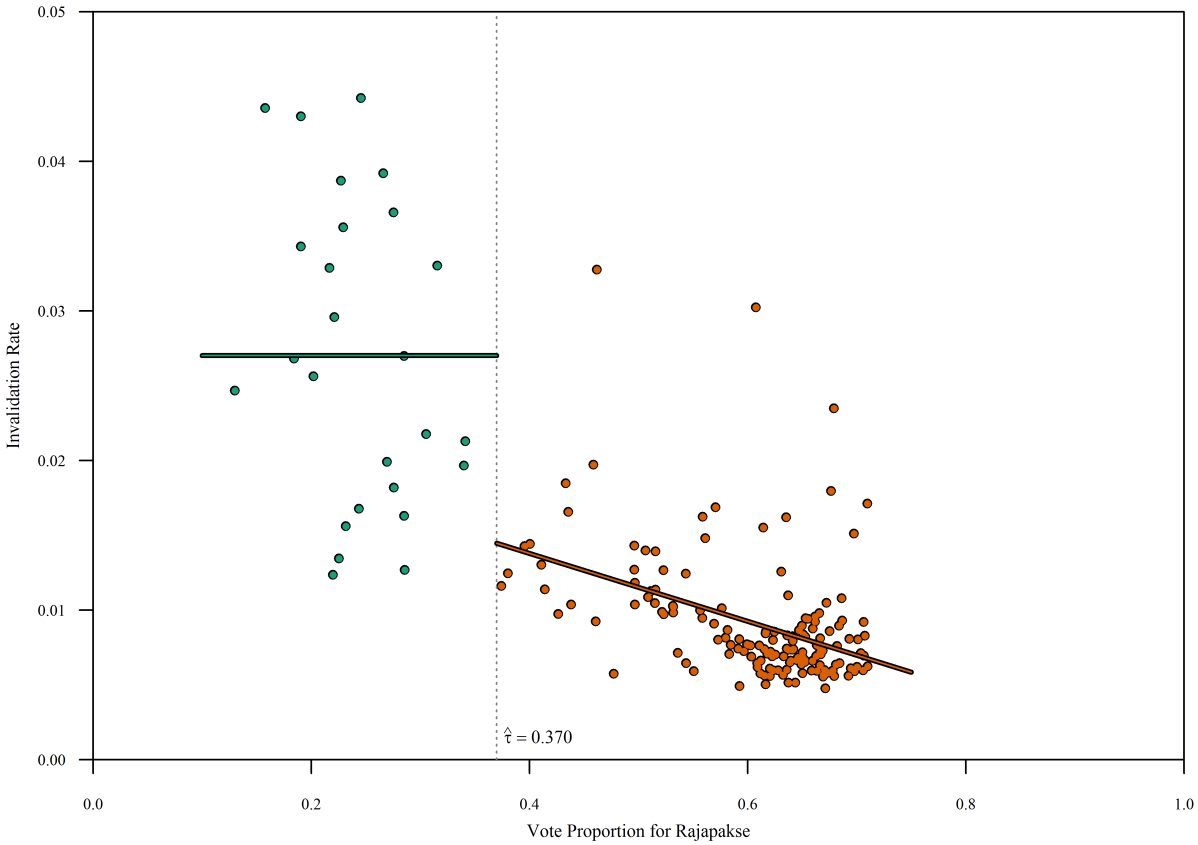


Figure 3.6: *Plot of the invalidation rate against vote support for President Mahinda Rajapaksa for the 2010 Sri Lankan presidential election. The threshold value of* $\tau = 0.37$ *arose solely from Figure 3.5.*

dation rates. Note that all thresholds produce segmented regressions of approximately equal quality.

I manually selected the threshold value of $\tau = 0.37$ as the greatest threshold before the sum of square errors began to consistently increase. Using that threshold value, the regression line on the points to the right has slope -0.023 ($p < 0.0001$). Figure 3.6 shows that the invalidation rate for those divisions with Rajapaksa support greater than $\tau = 0.37$ is not independent of Rajapaksa support. This is also very strong evidence against the free and fair hypothesis.

An Optimal Threshold: In the previous example, I manually selected the threshold by examining the graphic. A better method for estimating the threshold value τ is to

perform a modified CUSUM test. Page (1955) introduced the CUSUM test to detect a changepoint in a stationary time series. The CUSUM procedure consists of calculating the cumulative sum of the deviations from the mean of the series. The optimal changepoint is the point corresponding to the largest absolute cumulative sum. That is, let $\{y_i\}$ be a series of measurements taken at equal intervals. Calculate $C_j = \sum_{i=1}^{j} (\, y_i - \bar{y}\,)$ at each point in the series j. The optimal changepoint is

$$\tau^* = \operatorname*{argmax}_j |C_j|$$

However, this test is only able to detect a change in the mean of the series (Page 1955). To modify it to detect a change in the slope of the series, one merely has to perform the test on the first differences of the measurements. That is, let $\{y_i\}$ be a series of measurements taken at equal intervals. Define the first differences as $d_i = y_{i+1} - y_i$. Calculate $D_j = \sum_{i=1}^{j} \left(d_i - \bar{d} \right)$ at each point in the series j. The Page-optimal changepoint is

$$\tau_p^* = \operatorname*{argmax}_j |D_j|$$

Performing this procedure on the Ivoirian 2010 election provides an optimal threshold of $\tau_p^* = 0.067$. For the Sri Lankan 2010 election, the optimal threshold is $\tau_p^* = 0.470$. Performing the segmented regression test with this threshold produces similar results—the regression line on the electoral divisions to the right of the threshold has slope $\beta_1 = -0.016$ ($t = -3.19; p = 0.0017$). Again, this is strong evidence against the free and fair hypothesis.

Note that this analysis assumes the usual regression tests are appropriate under this selection procedure. Demonstrably, they are not. Figure 3.7 shows the distribution of the p-values of the usual WLS test of independence when the cutpoint is selected in this manner. The orange-colored bar represents the frequency of tests for which the calculated p-value was less than 0.05. If this is an appropriate test, we would expect the height of this bar to be approximately 5%. Here, the height is about 10%; that is,

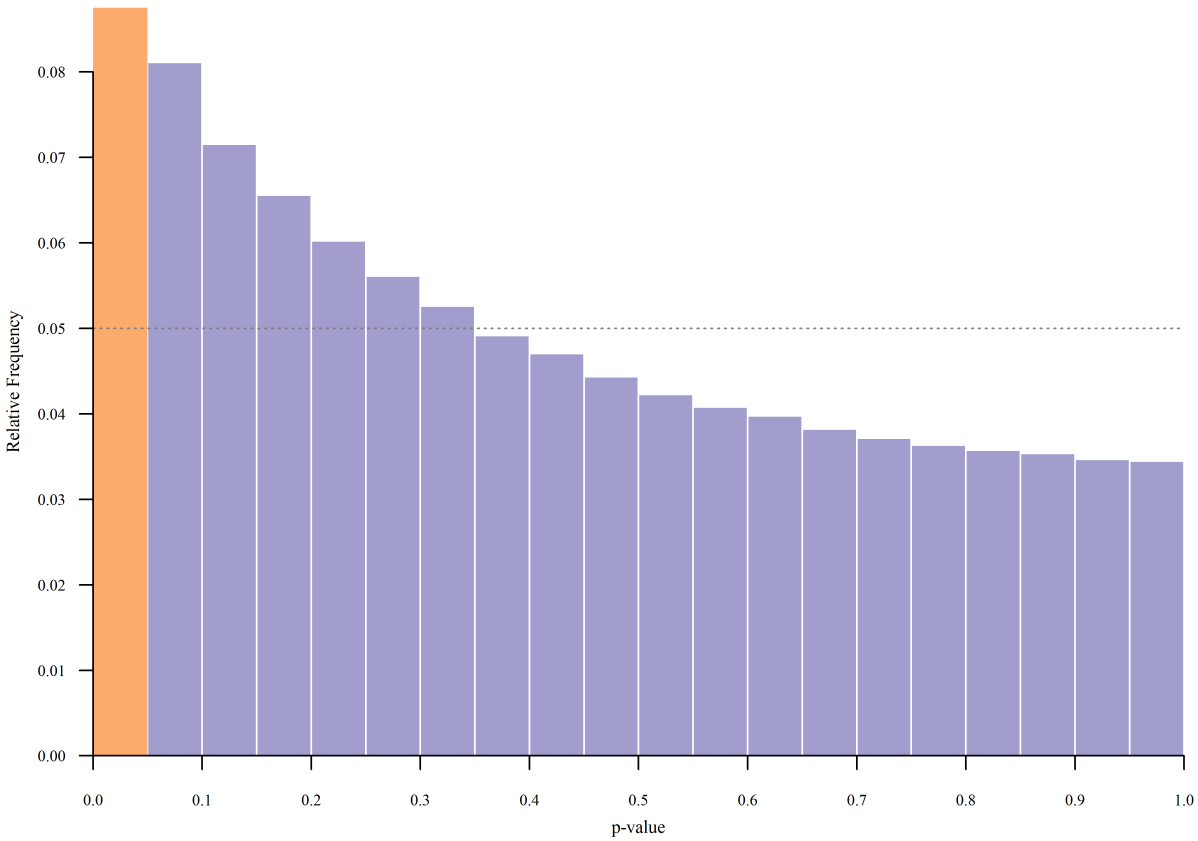


Figure 3.7: *Histogram of the calculated p-values based on the test described in the text. Note that the first bar, representing tests for which the p-value was less than 0.05, is about twice the expected height. The number of iterations performed was 1,000,000.*

using this procedure, one commits a Type I Error twice as often as expected. This is not a good feature of a statistical test. The distribution of p-values, for a continuous test statistic, should be standard Uniform (Westfall and Wolfinger 1997).

To adjust for this issue, three options present themselves. First, one may be able to derive the distribution of the test statistic from first principles. Second, one can use an estimated distribution for the distribution of the test statistic. Third, one is able to use Monte Carlo simulation to estimate the distribution of the test statistic—the parametric bootstrap. The first option may not be possible. The second option is of questionable utility. The third option can be done, but the calculations take time. However, if the simulations are run using the parameters from the election, namely the division sizes, the results are satisfactory and defensible.

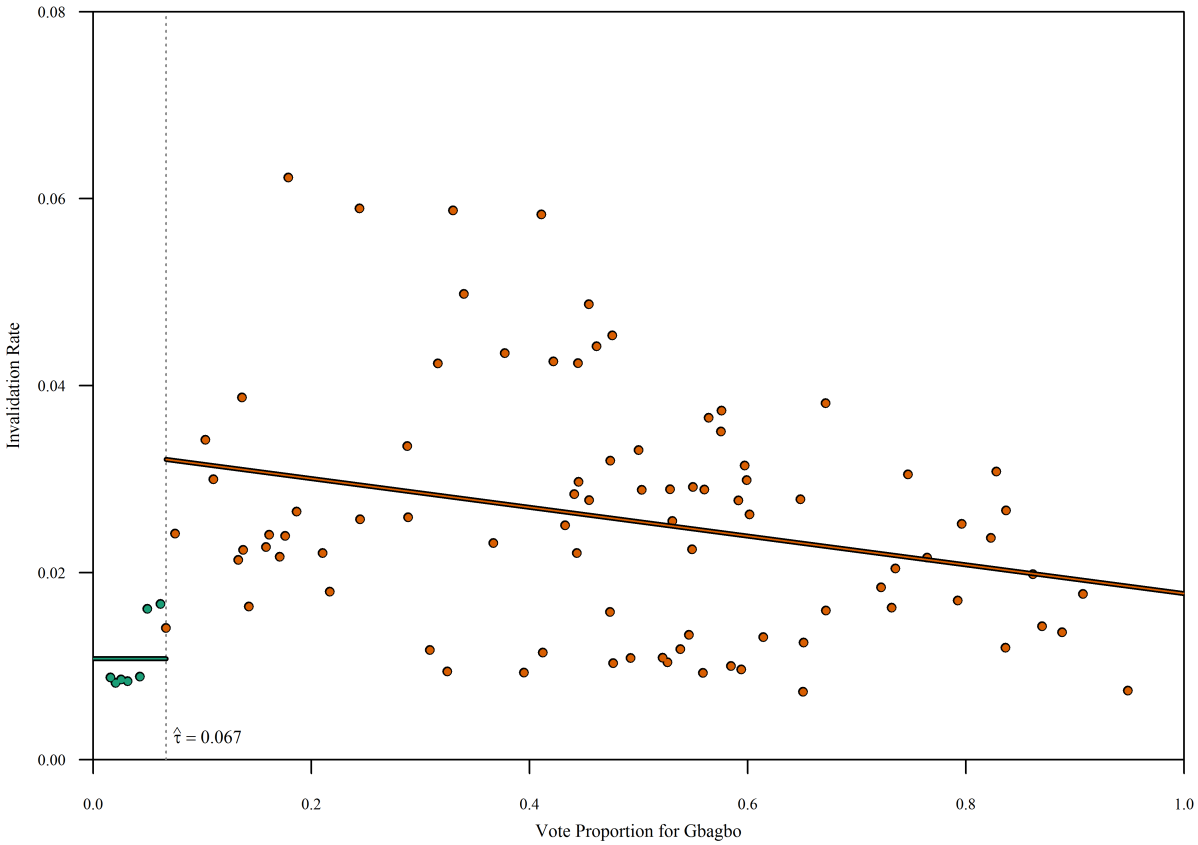


Figure 3.8: *Plot of the invalidation rate against vote support for President Laurent Gbagbo for the 2010 Ivoirian presidential runoff election. The optimal threshold value of* $\tau_p^* = 0.067$ *arose from Page's CUSUM test (1955).*

Using ten thousand simulations, the estimated critical values ($\alpha = 0.05$) for the Sri Lankan 2010 presidential election using simulation are -2.68 and 2.80. As the calculated test statistic for this election is -3.19—inside the rejection region. Furthermore, the estimated p-value is 0.0160, compared to the reported 0.0017. Thus, again, we have evidence that the Sri Lankan presidential election of 2010 lacks fairness, albeit not as strong as suggested by the usual test.

Returning to Côte d'Ivoire: Figure 3.8 provides the invalidation plot for the 2010 Ivoirian presidential runoff election. The Page-optimal threshold was estimated using the methods outlined above. The slope of the right segment is -0.015, with a t-value of -2.592. This is significantly different from zero at the usual $\alpha = 0.05$ level ($p = 0.0112$). From this, one would likely conclude that there is significant evidence of electoral fraud.

However, as I did with the Sri Lankan election above, I use simulation to estimate a corrected rejection region. Again, using 10,000 simulations, the estimated critical values are -2.94 and 3.04. This results in a simulated p-value of 0.098. Note that one cannot conclude, at the usual α level, that there is significant evidence of electoral fraud. This example shows the importance of not relying on the "usual" tests when the threshold is selected to optimize a quantity related to the model's goodness of fit.

3.3.2 Healy-Westmacott Regression. In this section, I introduce, explore, and use an Expectation-Maximization (EM) algorithm method developed by Healy and Westmacott (1956).

The EM algorithm is a general method designed by Dempster et al. (1977) to unify several related methods for estimating parameters in the presence of missing data. While Dempster et al. (1977) unified and expanded the methods, they did not create them. Two-score years earlier, Healy and Westmacott (1956) formulated a method for estimating parameters with missing data. While they worked in the realm of analysis of variance, their general method works well for regression. Regarding the missing data, Healy and Westmacott (1956, p 204) realized

> the fictitious values [estimated data] are actually the expected values of the missing units derived from the correct least-squares estimate

This method is an example of the EM algorithm when the likelihood is linear in the data (McLachlan and Krishnan 2008).

Let us again define our data model as

$$y_i = \alpha_0(x_i < \tau) + \beta_0(x_i \geq \tau) + \beta_1 x_i(x_i \geq \tau) + \varepsilon_i$$

The y_i and x_i are the invalidation rate and the candidate support in division i, respectively, and τ is the threshold between the two regimes.

Under this formulation, the Healy-Westmacott algorithm is

1. Fit the data with $\mathbf{Y}$, $\mathbf{X}$, and $\tau^{(k-1)}$. One can use any appropriate regression method. For reasons discussed in Section 3.2.3, this should be WLS or FGLS.

2. Calculate the value of $\tau^{(k)}$ as

$$\underset{\tau^{(k)}\in(0,1)}{\operatorname{argmin}} \left\{ \left\| \mathbf{Y} - \alpha_0 \left(\mathbf{X} < \tau^{(k)}\right) - \beta_0 \left(\mathbf{X} \geq \tau^{(k)}\right) - \beta_1 \mathbf{X} \left(\mathbf{X} \geq \tau^{(k)}\right) \right\| \right\}.$$

In this formulation, however, there is no "missing" data. The one parameter of interest, τ, can be estimated using only Step 2:

$$\tau^* = \underset{\tau\in(0,1)}{\operatorname{argmin}} \left\{ \left\| \mathbf{Y} \ - \alpha_0 \left(\mathbf{X} < \tau\right) - \beta_0 \left(\mathbf{X} \geq \tau\right) - \beta_1 \mathbf{X} \left(\mathbf{X} \geq \tau\right) \right\| \right\}.$$

This is its least squares estimate if $\|\cdot\|$ is the usual Euclidean norm.

The Healy-Westmacott algorithm, however, allows us greater flexibility: Perhaps there are two types of electoral divisions, but their dividing line is *not* a threshold on the candidate support rate. This algorithm allows us to separate all divisions into two groups, one with a zero slope and another with a (possibly) non-zero slope.

If we define $\mathbf{Z}$ as the vector of class membership, the Healy-Westmacott algorithm becomes

1. Fit the data with $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}^{(k-1)}$ to calculate $\alpha_0^{(k)}$, $\beta_0^{(k)}$, and $\beta_1^{(k)}$.

2. Calculate the value of $\mathbf{Z}^{(k)}$ as

$$Z_i^{(k)} = \begin{cases} 1 & y_i - \alpha_0 > y_i - \beta_0 - \beta_1 x_i \\ 0 & \text{Otherwise} \end{cases}$$

Step 2 assigns the division to the current closest regression line.

Note that this is a typical EM algorithm where the first step is to calculate the parameters based on the current data ($\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}^{(k-1)}$), and the second step is to update the estimated data ($\mathbf{Z}^{(k)}$) based on those parameters (McLachlan and Krishnan 2008).

Also note that this algorithm will produce two regression lines. Even under the null hypothesis, it is likely that the oblique regression line will have a statistically significant slope parameter. As such, we cannot rely only on that parameter. We must also take into consideration whether the two-line model is an improvement over the null model (a single horizontal line). There are asymptotic results when using the likelihood ratio statistic (Neyman and Pearson 1933), however this research is not about elections in Asymptopia. I strongly recommend using Monte Carlo to estimate the correct critical value(s).

As with the method of Section 3.3.1, the distribution of the regression test statistics is not Student's t. From preliminary exploration, it appears as though the actual distribution of the test statistic is stochastically less than that of the Student's t distribution; that is, the test will tend to reject at a rate in excess of α.

To handle this issue, the same three options present themselves. Again, I suggest estimating the p-value and confidence interval using simulation with experimental parameters equal to the parameters of the election.

To illustrate the results of this section, let us return to 2010 Sri Lanka. Figure 3.9 provides the invalidation plot. The green dots represent "fair" divisions; orange, "unfair." Note that the dividing line is not a vertical threshold.

To determine if this result is an improvement over the null model (all divisions green), I measure the ratio of the null model's squared standard error to that of the model graphed. The ratio is 1.67. To determine if this represents a significant improvement over the null model, I simulate 10,000 elections from the null distribution, collecting the test statistic for each. The upper 95th percentile is 1.57 and the estimated p-value is 0.0065. Thus, the two-line model is significantly better than the null model. The substantive conclusion is that we again have strong evidence against the free and fair hypothesis for the Sri Lankan 2010 presidential election.

Contrast this with the Ivoirian presidential runoff election. Figure 3.10 provides the usual invalidation plot with colors having the same meaning as above. The test

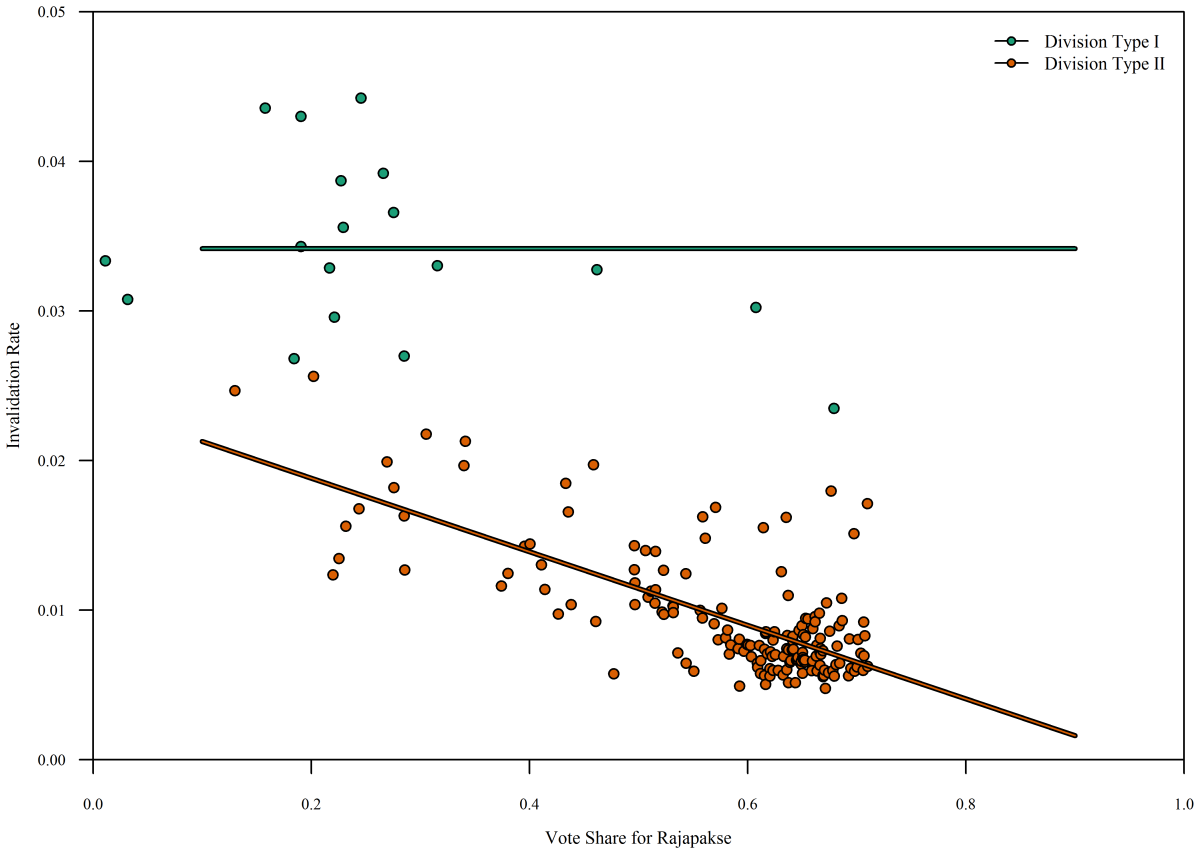


Figure 3.9: *Plot of the invalidation rate against vote support for President Rajapaksa for the 2010 Sri Lankan presidential election. The green divisions represent those following the null hypothesis; the orange divisions, the alternative.*

statistic (ratio of null residual variance to two-line residual variance) is 1.61. Estimating the distribution for this test statistic using 10,000 simulated (null) elections gives a critical value of 1.65, resulting in an estimated p-value of 0.080. Thus, we can conclude that this two-line model is *not* a significant improvement over the null model; that is, this test does not suggest evidence of electoral fraud in the election.

3.3.3 Empirical Bayes. The previous section did provide a generalization of the requirement that unfair divisions are those with candidate support above a specified threshold. However, it was ultimately unsatisfactory; it is rather difficult to believe that people in electoral divisions with very low support for the candidate are able to muster the ability to stuff the ballot boxes to a degree sufficient to affect the vote. Thus, while

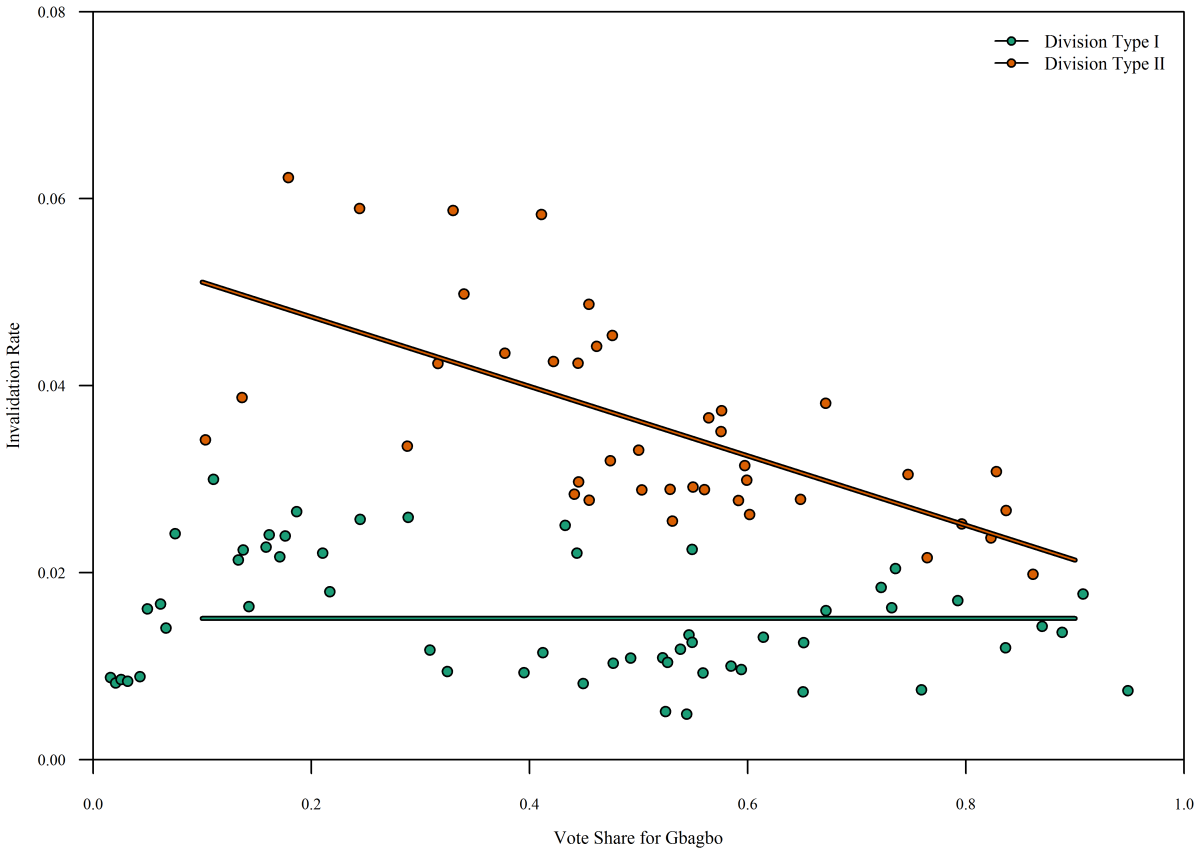


Figure 3.10: *Plot of the invalidation rate against vote support for President Gbagbo for the 2010 Ivoirian runoff presidential election. The green divisions represent those following the null hypothesis; the orange divisions, the alternative.*

that section *did* offer an interesting option to discovering two populations, the results have a tenuous connection to reality, at best. As such, let us return to estimating a threshold value separating fair and unfair divisions.

Note that in addition to the previous frequentist methods, one can use Bayesian methods. The advantage of Bayesian methods is that the entire distribution of all parameters of interest can be specified, under the assumption that the selected prior distributions are correct (Ntzoufras 2009).

Bayesian methods rely on Bayes' Law, which relates the prior distribution, the posterior distribution, and the likelihood of the data:

$$f(\theta \,|\, x) = \frac{f(x \,|\, \theta) f(\theta)}{f(x)}$$

Here, the posterior distribution of the parameters θ given the data is $f(\theta \mid x)$. The prior distribution of the parameters is $f(\theta)$, the data distribution is $f(x \mid \theta)$, and the probability of observing the data is $f(x)$. Fortunately, while $f(x)$ is usually unknown, it is a constant with respect to θ (Gelman et al. 2003). This leads to the "unnormalized posterior density"

$$f(\theta \,|\, x) \propto f(x \,|\, \theta) f(\theta)$$

The likelihood function arises from the data model. As before, the data model is

$$\mathbf{Y} = \alpha_0(\mathbf{X} < \tau) + \beta_0(\mathbf{X} \geq \tau) + \beta_1 \mathbf{X}(\mathbf{X} \geq \tau) + \boldsymbol{\varepsilon}$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(\mathbf{0},\ \sigma^2\mathbf{I}\right)$.

The prior distribution is selected by the researcher on the basis of prior information about the data-generating process. When the prior information is sparse, the prior distribution should be weakly-informative or non-informative. Multiple priors should be used to determine the sensitivity of the posterior distribution. Multivariate priors can be used, however products of univariate prior distributions tend to work "better" (Gelman et al. 2003; Jeffreys 1946).

Suggested Prior Distributions: In this model, there are five parameters for which priors need to be selected: τ, α_0, β_0, β_1, and σ^2. The following provides considerations for selecting appropriate priors for each.

The threshold value τ ranges from 0 to 1, exclusive. In the name of "letting the data speak," I suggest selecting a low-information prior distribution, specifically

$$\tau \sim \text{BETA}(1, 1)$$

This is the Uniform distribution.

The constant term for the below-threshold points α_0 also ranges from 0 to 1, exclusive. The average invalidation rate makes a natural expected value for α_0. I suggest

selecting a data-dependent prior distribution of low information:

$$\alpha_0 \sim \text{BETA}\left(\frac{\bar{y}}{1-\bar{y}}, 1\right)$$

According to this distribution, the expected value of α_0 is $\mathbb{E}[\alpha_0] = \bar{y}$.

Along the same vein of letting the data speak, I suggest low-information prior distributions of

$$\begin{aligned}\beta_0 &\sim \mathcal{N}(0, 100)\\ \beta_1 &\sim \mathcal{N}(0, 100)\end{aligned}$$

for these parameters.

Finally, the variance of $\boldsymbol{\varepsilon}$ is a latent variable whose distribution only needs to be sufficiently non-informative and positive. For this reason, I suggest $\sigma^2 \sim \text{GAMMA}(0.01, 0.01)$.

While these distributions are consistent with the problem, variations should also be used to demonstrate the sensitivity of the results on the prior distribution selections.

Gibbs Sampling: With the selection of the prior distributions, the posterior distributions need to be estimated. While several techniques exist for this estimation, Gibbs Sampling is popular (Geman and Geman 1984). It is based on the full conditional distributions of the parameters. The algorithm starts with an initial value for all of the parameters, $\theta_i^{(0)}$, for all i.

A random value is generated for each of the parameters based on the full conditional distributions, $p(\theta_i \mid \theta_{-i})$, the distribution of θ_i given the value of all other parameters. A second random value is generated for each of the parameters based on the same full conditional distributions. These iterations continue until the researcher determines the estimated parameters have converged in distribution to their true (target) distribution (Albert 2009).

Unfortunately, such convergence is difficult to determine *a priori*. One rule of thumb is to examine sequential plots of each parameter for evidence of complete mixing. A second method is to compare parameter distributions at several places along the chain.

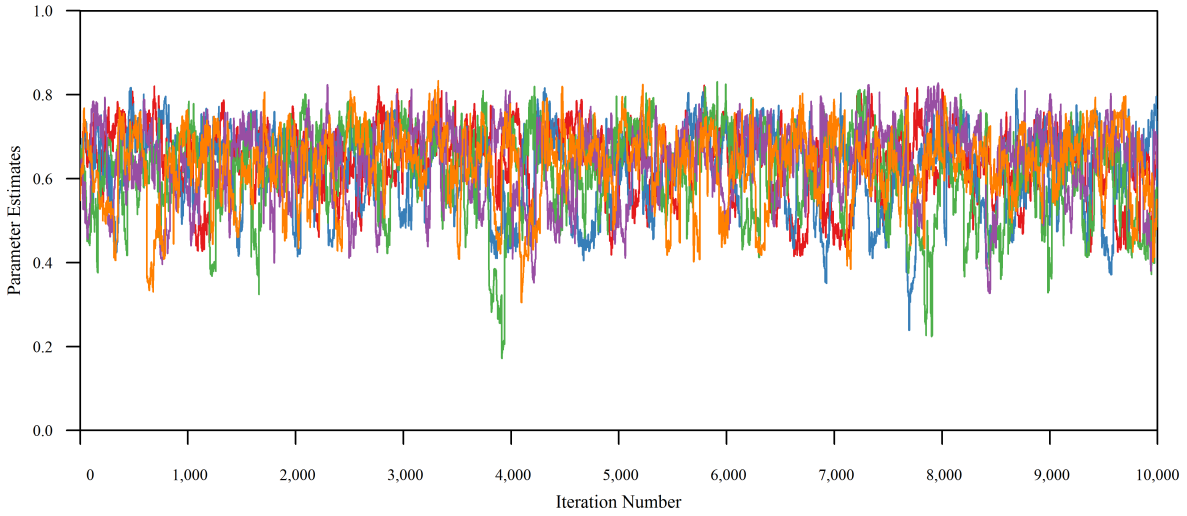


Figure 3.11: *A sequential plot (index plot) of parameter estimates for an example parameter using the Gibbs Sampling algorithm. There are five chains. Note the evidence for mixing rather quickly. Based on this plot, I would select 1000 as the burn-in period.*

A third method is to start several chains in different places and look to see when they are similar enough (Gelman et al. 2003; Ntzoufras 2009).

To illustrate this, I created a simulated election with a clear changepoint ($\tau = 0.600$) and very little noise in the dependent variable:

```
x = runif(n, min=0.1,max=0.9)
r = rep(0.05, n)
r[x>tau] = -0.1*x[x>tau] + 0.11
y = r + rnorm(n, m=0,s=0.0001)
```

For this election, Figure 3.11 shows plots of five chains. Note that, beyond the first few iterations, there does not appear to be much difference in the five chain distributions. From this, one could conclude that the burn-in period is small, perhaps 1000.

In addition to determining a burn-in period and a collection period, the researcher must ensure that there is no serial correlation in the parameter estimates. If such exists, the parameter estimates should be 'thinned' to reduce the correlation (Ntzoufras 2009). There is *very* strong autocorrelation in all five chains. This is troubling.

Once the target distribution is reached, one has estimates for all interesting parameters in model. These estimates include the mean, median, percentiles, etc.

Parameter	Mean	StdDev.	$P_{0.025}$	$P_{0.50}$	$P_{0.975}$
α_0	0.04961	0.00182	0.04614	0.04958	0.05325
β_0	0.08224	0.03484	-0.00671	0.08837	0.13770
β_1	-0.06520	0.04293	-0.13490	-0.07168	0.04282
τ	0.62809	0.09312	0.42710	0.64550	0.76770

Table 3.2: *Summary statistics for the posterior distributions estimated for the simulated election with the prior distributions provided in the text.*

An Illustration: To illustrate the process, let us return to the generated election described previously. The true parameter values are $\alpha_0 = 0.05$, $\beta_0 = 0.11$, $\beta_1 = -0.10$, and $\tau = 0.60$.

The data were fit using this Bayesian procedure. The burn-in was 1000, the chain length was 10,000, and the number of chains was 5. Table 3.2 provides the parameter estimates. Note that for this toy example, the Bayesian procedure recovered the parameter estimates, most importantly the threshold. For estimation, I used OpenBUGS with the `R2OpenBUGS` library (Lunn et al. 2009; Sturtz et al. 2005).

The parameter of most interest in this setting is τ, which is the optimal threshold between the fair and the unfair elections. Table 3.2 provides its summary statistics. The 95% Bayesian credible interval is from 0.43 to 0.77, with a mean of 0.63 and median of 0.65. The true value is 0.60, which is within the interval.

Plotting the effects is also instructive. Figure 3.12 plots each simulated electoral division. The Bayesian-optimal threshold (posterior mean) is shown. Divisions are colored based on their position with respect to that threshold. The two regression lines are plotted.

Note that while the method estimates the threshold τ well, it may not estimate the other parameters well. In particular, the slope β_1 in the right hand region is particularly incorrect.

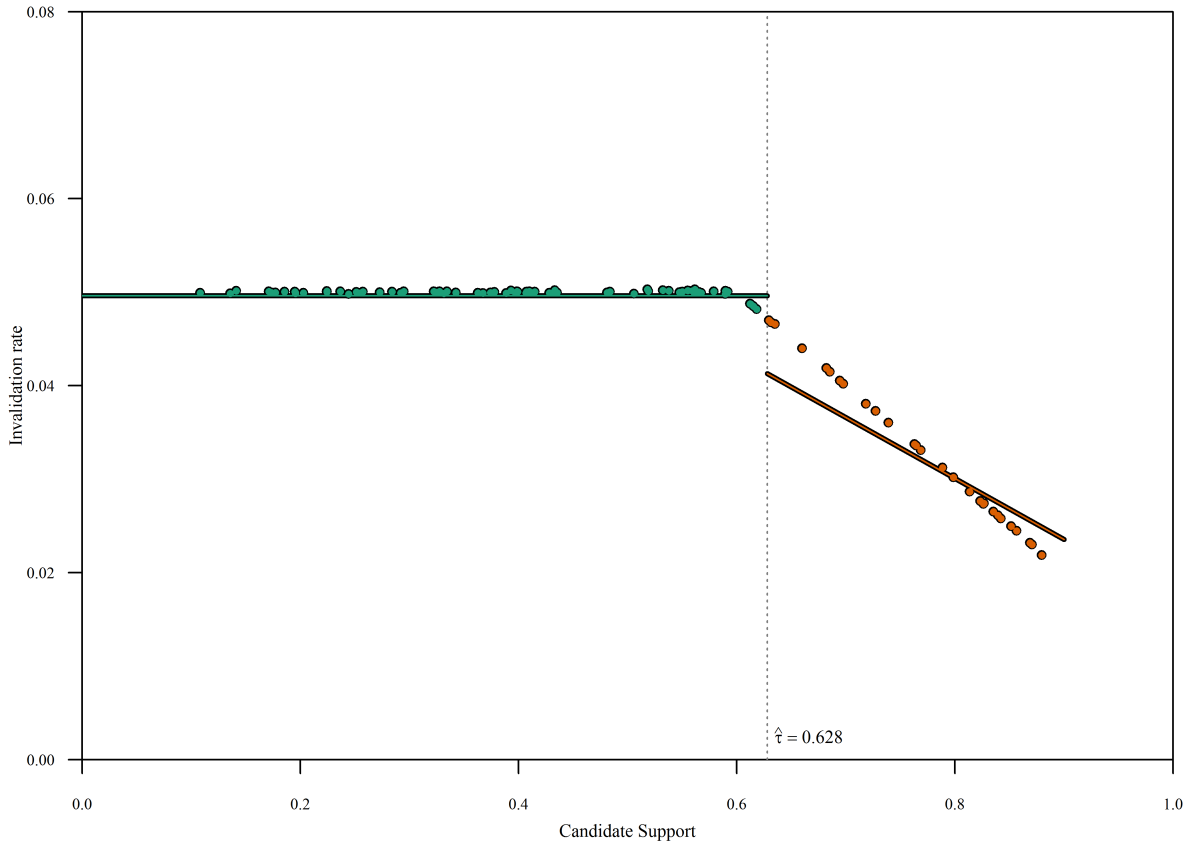


Figure 3.12: *Scatterplot of the invalidation rate against the support for the simulated election discussed in the text. Posterior means were used for parameters of the regression lines.*

3.3.4 A Comparison of These Methods. In this part of the chapter, I covered three methods for meaningfully partitioning the data. The first method is a simple grid search maximizing the mean square error. The second method is the Healy-Westmacott algorithm. The final method is Bayesian regression. For reasons discussed at the start of Section 3.3.3, I withdrew the Healy-Westmacott estimator from consideration.

To determine which method is preferred, I will use the mean square error for the reasons discussed in Section 1.3. I selected this metric as it combines both the bias of the estimator and its variability.

$$\text{MSE} = \text{bias}^2 + \sigma^2$$

Noise Level	Grid Search	Bayesian Posterior Mean	Bayesian Posterior Median
0.0010	0.000 259	0.001 653	0.000 402
0.0056	0.000 274	0.002 141	0.000 475
0.0316	0.000 507	0.001 909	0.000 449
0.1778	0.002 494	0.004 153	0.001 997
1.0000	0.041 021	0.044 498	0.069 044

Table 3.3: *The mean square errors for the three threshold estimation methods discussed in this section. The number of elections analyzed is 10,000 for the grid search, but only 1000 for the two Bayesian methods.*

The Simulated Elections: First, I needed to create several elections. Two variables were created for each election, an invalidation rate and a candidate support rate. Elections with candidate support rate less than the threshold were created fair; the two variables were independent. Elections with candidate support rate greater than the threshold were created unfair; the two variables were dependent.

The basic code I used to generate the elections is given on page 86. I set the threshold to $\tau = 0.600$, α_0 to 0.05, n to 100, β_0 to 0.11, and β_1 to -0.10. The noise level is the standard deviation of the ϵ variable. I generated 10,000 elections for five different noise levels from 0.001 to 1.000. This range allowed me to determine which of the methods best estimated the threshold value at different noise levels. When the noise level is small, the relationship and threshold are very evident (e.g., Figure 3.12). When it is large, there is neither an apparent relationship nor an apparent cutpoint.

The grid search algorithm ran relatively fast, analyzing 10,000 elections in approximately an hour. The Bayesian method, however, was much slower. Each election took approximately 1.9 seconds to analyze. For this reason, I only analyzed 1000 elections for each of the five noise levels using the Bayesian technique.

Table 3.3 lists the mean square errors for each of the methods. Note that the Bayesian methods consistently have the greatest mean square error of the group. In terms of execution times, the Bayesian methods are significantly slower than the grid

search. The only weakness with the grid search was that it did not always produce an appropriate estimate. When the noise level was 1.000, the grid search failed to produce a threshold estimate 167 times out of the $10,000$ elections; that is, in 167 elections, the grid search estimated τ at the edge of the parameter space. The Bayesian methods always produced an estimate.

3.4. Conclusion

The previous chapter examined several techniques for testing the free and fair hypothesis for vote counts. In this chapter, I covered several techniques one can use when the invalidation rates are also available. These techniques centered on regression tests.

Those regression tests blended two aspects. First, I noted that it is likely that electoral divisions are of two populations: fair and unfair. Current regression use in electoral forensics does not leverage this information, it assumes a single population. In doing so, current tests are not as powerful as they could be.

Second, I noted that the use of ordinary least squares regression may be inappropriate for this type of data; different electoral divisions have different weights associated with them (*cf* Section 7.1). These weights are due to the very number of ballots cast. When the division has a low (absolute) turnout, it should lend less information to the overall regression. When the division has a high turnout, it should lend much more information to the regression. Ordinary least squares (OLS) ignores this additional information.

To solve the issue of division sizes, I suggested feasible generalized least squares (FGLS). This method iteratively estimates the unknown covariance matrix. In doing so, FGLS reduces the bias in the standard error estimates. Note that FGLS is an iterative procedure; the first iteration is an OLS estimation, the following are weighted least squares (WLS), with adjustments to the covariance matrix at each iteration.

Feasible GLS produces estimates very similar to those of WLS when the heteroskedasticity is not a function of an independent variable. As the source of the heteroskedasticity in this research is the district vote count and not the vote for the candidate, the results of FGLS are very close to those of WLS. As WLS is faster than FGLS, I recommended weighted least squares.

The solution to the two-populations issue came from assuming that there was a threshold present in the election. I assumed that those divisions with a candidate support rate less than that threshold were fair; those above, unfair.

While these assumptions may not be strictly met in reality, they will be good approximations, especially when the probability of electoral fraud is dependent on the candidate's support in the division.

To estimate the threshold, I proposed three general methods. The first was a simple grid search. The second was a variation on the Expectation-Maximization algorithm proposed by Healy and Westmacott. The final was a Bayesian model. Of these three, the Healy-Westmacott algorithm proved itself poor in the context of electoral forensics.

I evaluated the remaining two methods with respect to three factors: mean square error, execution time, and estimate production. By the first two metrics, the grid search was superior to the Bayesian method. However, the grid search infrequently failed to produce estimates.

Thus, we close this chapter with the following solution. When the government publishes vote counts *as well as* invalidation rates, one should use the grid search algorithm to estimate the threshold. Once the threshold is determined, weighted least squares should be used to test the free and fair hypothesis: $H_0 : \alpha_0 = \beta_0$ and $\beta_1 = 0$. One must also remember to use simulation to estimate the critical values and p-values.

In the next chapter, I emphasize that voting is a spatial process. I then explore how this affects the previous tests. I conclude the chapter by comparing the current geographically weighted regression and my proposed spatial lag expansion model.

The abbreviated R code for conducting feasible generalized least squares:

```
fgls.abbr <- function( x,y,N,
    tol.est=1e-6,
    tol.var=tol.est,
    max.iter=10) {

 n = length(x)

## Make the matrices
 X    = matrix( c(rep(1,n),x), ncol=2 )
 Y    = matrix(y, ncol=1)
 M    = diag(N)
 Minv = diag(1/N)

## Get initial estimates (OLS Step)
 bHat = solve(t(X)%*%X) %*% t(X)%*%Y
 pHat = bHat[1]
 resd = as.numeric(Y-X%*%bHat)

 sHat = as.numeric(t(Y-X%*%bHat)%*%M%*%(Y-X%*%bHat) /( (n-2)*pHat*(1-pHat
   ) ))
 vHat = diag(1/N)*pHat*(1-pHat)*sHat

## Get iterated estimates (FGLS Steps)
 for( i in 1:max.iter ) {
  bHat.last=bHat
  sHat.last=sHat
  vHatInv = solve(vHat)
  bHat = solve( t(X)%*%vHatInv%*%X) %*% t(X)%*%vHatInv %*% Y

  if( sum((bHat.last-bHat)^2)<tol.est )  break

  pHat = bHat[1]
  sHat = 1/(n-2) * t(Y-X%*%bHat) %*% M %*% (Y-X%*%bHat) / (pHat*(1-pHat))
  vHat = as.numeric( sHat * pHat *(1-pHat) ) * Minv

  if( abs(sHat.last-sHat)<tol.var )  break

 }

## Estimate the estimate variance
 sEst = solve(t(X)%*%solve(vHat)%*%X)

## Prepare the output
 se = sqrt( sEst[2,2] )
 p.val = pt( -bHat[2] / sEst[2,2]^(0.5), df=n-2 )
 res = 1 - 2*abs(0.5-p.val)

return(res)
}
```

CHAPTER 4

CONSIDERING GEOGRAPHY

Sri Lanka, 2010. In 2009, the civil war in Sri Lanka finally ended. The rebel Tamil group in the north of the island state finally succumbed to the overwhelming drive of the south to keep the island united under a Sinhalese ethnicity—with input from the Tamil minority. This war, which saw the deaths of thousands and the introduction of strategic suicide terrorism, ended with a combination of military acumen and natural disasters—the 2004 Boxing Day tsunami ravaged the eastern part of the island, where the rebel Liberation Tigers of Tamil Eelam (LTTE) had stationed their navy. President Mahinda Rajapaksa, as the titular head of the military, and General Sarath Fonseka, as operational head of the military, led the final victory (Forsberg 2012).

The two were close friends through and until the end of the war, a fact that may have led to the military ultimately being successful. However, the two public figures had a falling out, and Fonseka entered the political fray in 2009, campaigning in the January 2010 presidential election against his former friend and ally.

Polls showed a very close race, which was to be expected, as both were held in high esteem by many in Sri Lanka. Because of that, the British Broadcasting Corporation (BBC) predicted that the initial results may not show a clear winner and that it would take upwards of a week to know the winner (BBC News 2010).

The run-up to voting day had the typical Sri Lankan violence: several died in clashes between the supporters of Rajapaksa and others. Election day itself progressed with the expected intimidation and violence at polling stations (CMEV 2010). However, since no international observers were allowed to observe the election, the only evidence

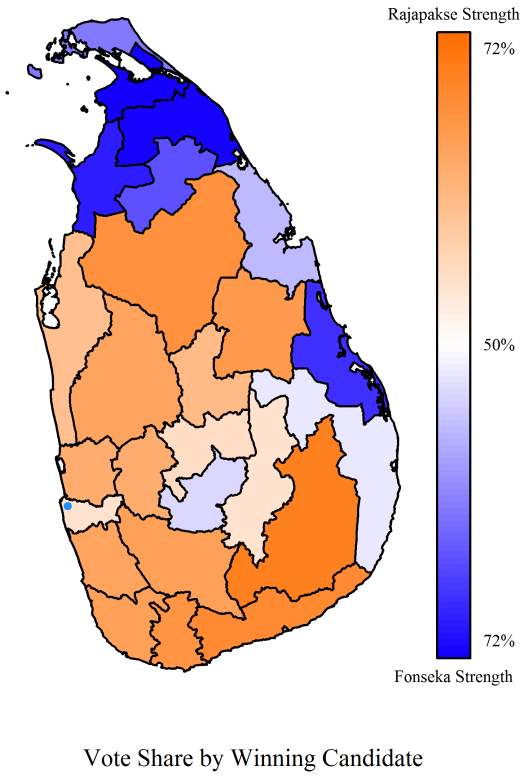


Figure 4.1: *Map showing the level of candidate support for the 2010 Sri Lankan presidential election. Support for incumbent President Mahinda Rajapaksa is shown in orange; challenger General Sarath Fonseka, in blue. Darker colors indicate higher support. The capital is marked with a blue dot.*

of these events were the cold bodies and the conflicting claims by rival factions (BBC News 2010).

The Sri Lankan Department of Elections began counting the ballots on schedule, and the election winner was known within a few hours. The election race that was supposed to be close ended with Fonseka officially receiving 40% of the vote, Rajapaksa, 58% (CMEV 2010).

A map of the support for the winning candidate is provided in Figure 4.1 (GADM 2014c; Sri Lanka 2010). Rajapaksa support is indicated in orange; Fonseka, in blue. Note that Fonseka did well in the Tamil north and east; Rajapaksa in the Sinhalese south and west. Rajapaksa carried the capital of Sri Jayawardenepura Kotte in the west (Colombo

District, Western Province), marked with a blue dot. Fonseka carried his home Nuwara Eliya District (Central Province).

The next day, Fonseka declared the results invalid due to voting irregularities. At the four-star Cinnamon Lakeside hotel, Fonseka gathered his closest advisors to decide his course of action. Later that morning, Rajapaksa ordered the hotel surrounded by the military, effectively separating Fonseka from the outside world. Rajapaksa first accused Fonseka of plotting a military coup, then of military offenses, and then of running for president while in the military (BBC News 2010).

Through it all, Fonseka claimed widespread electoral fraud. The final report on the election from the independent Center for Monitoring Election Violence agrees (CMEV 2010). The response from the US Embassy in Sri Lanka is silent on the issue (US State Department 2010):

> The United States congratulates Sri Lanka for the first nationwide election in decades and President Rajapaksa on his victory. We look forward to continuing the partnership between our two countries and working with the Government and the people to support a peaceful and prosperous Sri Lanka.

Today, Mahinda Rajapaksa is the President of Sri Lanka. General Fonseka was found guilty of treason and sentenced to 20 years. Rajapaksa had Fonseka released on May 21, 2012 (BBC News 2012).

4.1. Introduction

The previous chapter covered several regression tests to detect violations of the free and fair hypothesis, specifically the assumption that the invalidation rate is independent of the candidate support rate. When these two variables are not independent, there is *prima facie* evidence of electoral fraud.

However, those tests can be also applied to testing for inequalities in the electoral system, and not just fraud. Regressing the invalidation rate on the proportion of elderly in the electoral division can test if the system is biased against the elderly. Regressing the invalidation rate on the proportion of minorities can test if the system is biased against minorities. Furthermore, it would be most appropriate to use several independent variables at once. This will allow one to better test the independence hypothesis while controlling for several variables (Chapter 6).

The previous tests tend to ignore the fact that voting is an inherently spatial event. People near each other will tend to vote similarly (Mebane and Sekhon 2004). As such, ignoring the geographical component omits relevant information. First, the residuals may be spatially correlated. This would be a violation of most techniques discussed in Chapter 3. Second, the effects may vary across the country. Not only would this be a violation of the assumptions of all previous techniques, it would be interesting in itself.

Figure 4.2 illustrates that spatial correlation seems to be an issue with some elections. The left map shows the invalidation rate in each Sri Lankan province. Darker orange colors correspond to higher invalidation rates. The map suggests a hot spot for invalidation in the north of the island. The right map shows the proportion of the vote cast for incumbent President Mahinda Rajapaksa. The map also suggests a cold spot in the north. Is the north different in both respects from the rest of Sri Lanka? That is an interesting question.

This chapter concerns geography. First, I cover three current methods for modeling geographic information. Those three methods are lagged variable regression, expansion method, and geographically weighted regression. Second, I introduce a new method that improves upon the first two and corrects some of the issues with the third. Before I tackle these topics, I discuss detecting spatial correlation.

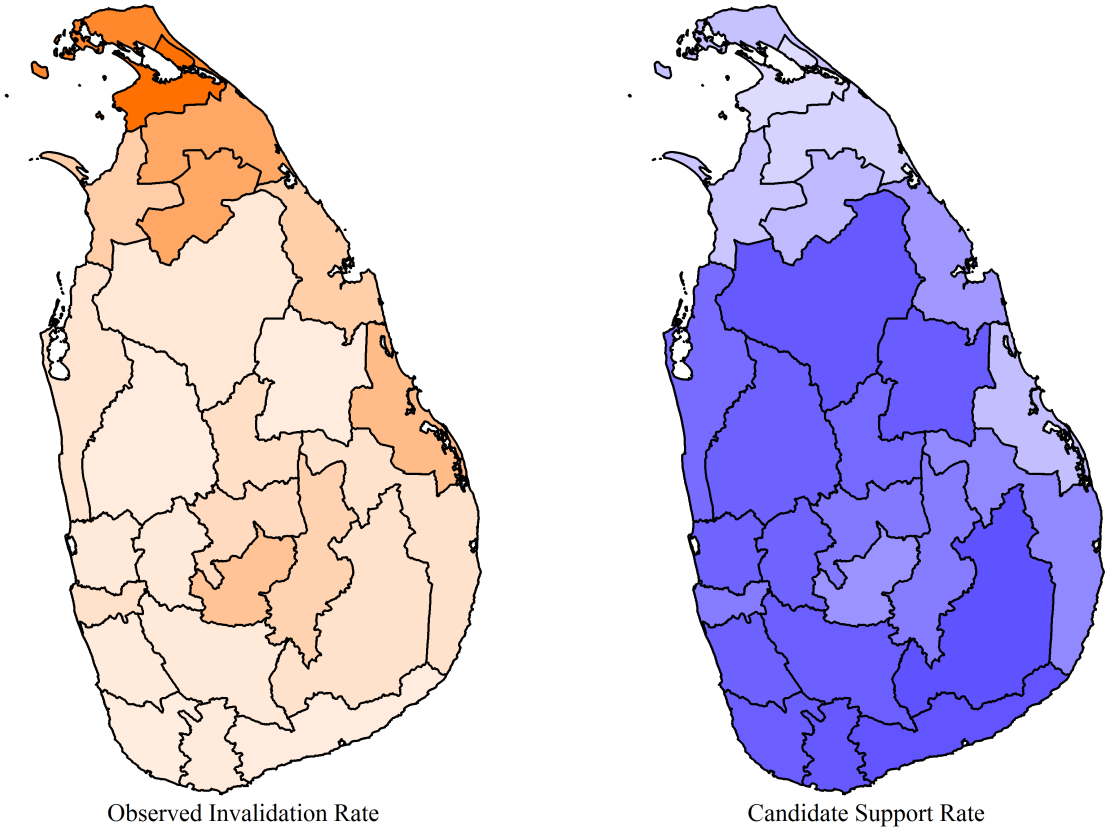


Figure 4.2: *Maps showing the invalidation rate(left) and the level of support for President Rajapaksa in the 2011 Sri Lankan presidential election.*

4.2. Detecting Spatial Correlation

Spatial correlation is the situation in which the value at a point is correlated with values at other points. As with other types of correlation, spatial correlation can be positive or negative. If the data have positive spatial correlation, then nearby values will be similar. If the data have negative spatial correlation, then nearby values will be very dissimilar. Figure 4.3 shows examples of negative (left), zero (center), and positive (right) spatial correlation. Note that the spatial correlation is measured using rook continuity in all three cases; that is, cells are neighbors only if they abut east-west or north-south, not if they are diagonal.

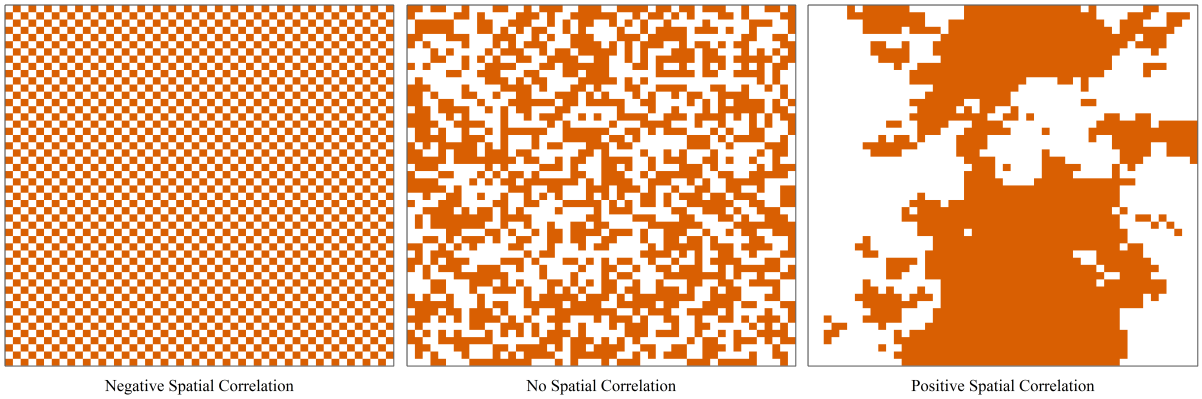


Figure 4.3: *Three graphics demonstrating spatial correlation. The left panel shows strong negative spatial correlation; the center, none; the right, high positive. In all three cases, neighbors are measured using rook contiguity; that is, neighbors are horizontal or vertical, not diagonal.*

4.2.1 The Neighbor Matrix. That brings up the first topic: Neighbors and the neighbor matrix (also known as the adjacency or the contiguity matrix) **W**. The purpose of the **W** matrix is to specify the expected pattern of geographical correlation. The entries in the neighbor matrix consist of values in $[0, 1]$. These values indicate which entities are neighbors, and to what extent. Figure 4.4 provides a simple map with eight entities, labeled A through H, with measures taken on each entity.

Contiguity type follows chess terms and meanings. According to the rook contiguity, the following ten pairs in Figure 4.4 are neighbors: (A,B), (A,E), (B,C), (B,F), (C,D), (C,G), (D,H), (E,F), (F,G), and (G,H). Under bishop contiguity, the following six pairs are neighbors: (A,F), (B,E), (B,G), (C,F), (C,H), and (D,G). As the queen can move as a rook or as a bishop, the queen-contiguous neighbors are the union of the rook- and bishop-contiguous: (A,B), (A,E), (B,C), (B,F), (C,D), (C,G), (D,H), (E,F), (F,G), (G,H), (A,F), (B,E), (B,G), (C,F), (C,H), and (D,G).

While this example described neighbors as touching, there are extensions to the rook-, bishop-, and queen-contiguities. Those described above are first-order contiguous. Second-order contiguous includes neighbors of neighbors; third-order includes neighbors of neighbors of neighbors; and so forth.

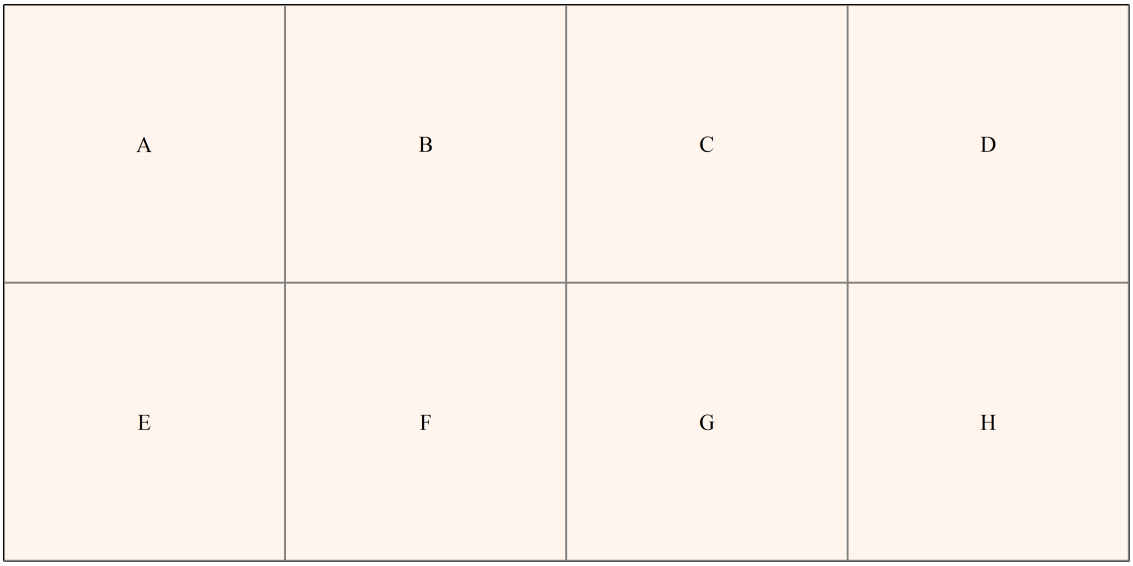


Figure 4.4: *A simple* 2×4 *map to illustrate the concepts discussed in the text.*

The map of Figure 4.4 shows that all entities are at most third-order queen-contiguous, or fourth-order rook-contiguous. A shortest path from A to H is A - B - C - H using queen contiguity, is A - B - C - D - H using rook contiguity, and is A - F - C - H using bishop. Note that A and E are not neighbors of any order under bishop contiguity.

With this description, the neighbor matrix for first-order rook contiguity, assuming you are your own neighbor (a.k.a. zeroeth-order neighbor), is

$$\mathbf{W} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} \tag{4.1}$$

Each i, j entry indicates whether cell i and cell j are contiguous according to the rook definition.

The cells need not indicate *binary* contiguity. Another type of neighbor matrix includes an inverse distance as the cell entries. This distance can be Euclidean, geodetic, or some other measure.

This discussion of contiguity is important as it determines to a large part whether spatial correlation can be detected. For instance, the left panel of Figure 4.3 shows highly negative spatial correlation using rook contiguity. However, were we to use bishop contiguity, it would have strong positive spatial correlation. Furthermore, were we to use queen contiguity, the spatial correlation would be close to zero.

4.2.2 Measures of Local Spatial Correlation. Now that we have defined neighbors, detecting spatial correlation is relatively easy. There are several measures of local spatial correlation. These measure differ from the global measures in that they calculate correlation at each point, rather than the map as a whole.

There are several available measures. Moran's I_i (Anselin 1995), Geary's C_i (Geary 1954), Getis and Ord's G_i^* (Getis and Ord 1997), and Hatfield's H_i (Hatfield 2011) all measure different aspects of local spatial correlation. In lieu of using all measures, I use the popular Moran's Local I_i (Anselin 1995).

Definition 4.1 (Moran's Local I_i Statistic)**.** *Let w_{ij} be the $(i,j)^{th}$ entry in the neighborhood matrix with $w_{ii} = 0$. Define $z_i = x_i - \bar{x}$, where x_i is the measurement at the i^{th} location and $\bar{x}$ is the average of those measurements across the neighbors of i. Moran's Local I_i statistic is*

$$I_i := \frac{z_i}{m_2} \sum_{j=1}^{n} w_{ij} z_j$$

Here, $m_2 = \frac{1}{n} \sum_{i=1}^{n} z_i^2$.

A usual next step is to determine the distribution of the I_i statistic. To do so, we need to make assumptions about the data values. The usual assumption here is the "randomization assumption." This assumption holds the measured values fixed (non-stochastic), but randomize the entities to which they belong. Among other things, this

makes the sample second moment, m_2, non-stochastic. Thus, under the randomization assumption, we have the following result (Anselin 1995).

Proposition 4.2 (The expected value of Moran's I_i)**.** *Let* $X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. *Also, as above, define* $z_i = x_i - \bar{x}$. *If* $w_i := \sum_j w_{ij}$, *then* $\mathbb{E}[I_i] \approx -\frac{w_i}{n-1}$.

Proof. This proof follows from definitions and algebra.

$$\begin{aligned}\mathbb{E}\big[I_i\big] &= \mathbb{E}\left[\frac{Z_i}{m_2}\sum_j w_{ij}Z_j\right] \\ &= \sum_j w_{ij}\mathbb{E}\left[\frac{Z_iZ_j}{m_2}\right]\end{aligned}$$

Under the randomization hypothesis, m_2 is fixed. Furthermore,

$$\begin{aligned}\mathbb{E}\big[Z_iZ_j\big] &= \frac{1}{n-1}\mathbb{E}\left[Z_i(\sum_{j\neq i} Z_j)\right] \\ &= \frac{1}{n-1}\mathbb{E}\left[Z_i(\sum_{j} Z_j - Z_i)\right] \\ &= -\frac{1}{n-1}\mathbb{E}\big[Z_i^2\big] \\ &= -\frac{1}{n-1}m_2.\end{aligned}$$

Back substitution gives our result: $\mathbb{E}[I_i] = -\frac{1}{n-1}w_i$. □

Under the same randomization assumption, and following Cliff and Ord (1981), Anselin (1995, Page 99) also shows that the variance of Moran's I_i is the unwieldy

$$\mathbb{V}\big[I_i\big] = \frac{w_{i(2)}(n-b_2)}{n-1} + 2w_{i(kh)}\frac{2b_2-n}{(n-1)(n-2)} - \frac{w_i^2}{(n-1)^2}$$

Here, $w_i = \sum_j w_{ij}$, $b_2 = m_4/m_2^2$, the second moment $m_2 = \sum_i z_i^2/n$, the fourth moment $m_4 = \sum_i z_i^4/n$, $w_{i(2)} = \sum_{j\neq i} w_{ij}^2$, and $2w_{i(kh)} = \sum_{k\neq i}\sum_{h\neq i} w_{ik}w_{ih}$.

As this variance is finite, the Central Limit Theorem tells us that the distribution of I_i converges in law to the Normal distribution. With this, one can perform the usual hypothesis tests and can calculate the usual confidence intervals.

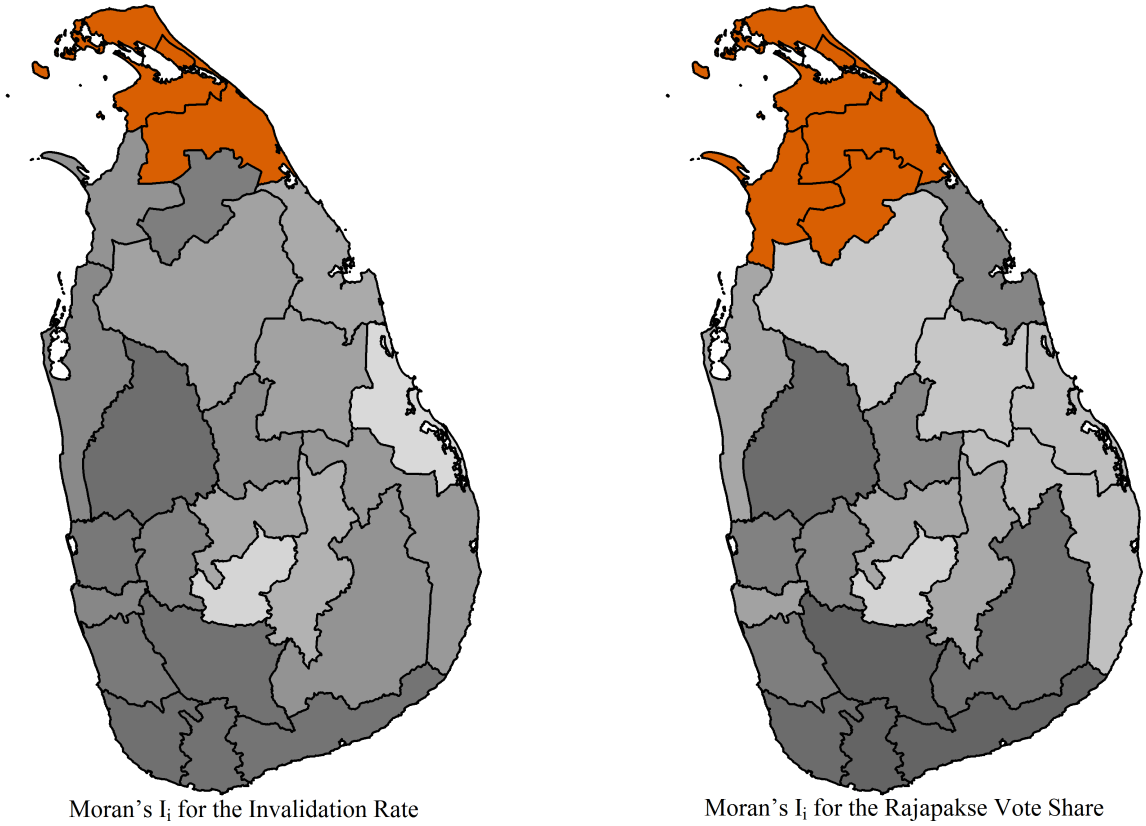


Figure 4.5: *Maps of Moran's local I_i for the Sri Lankan presidential election, 2010. The left panel shows the p-value for Moran's measure for each of the 25 districts. The right panel shows the same for the vote proportion for Rajapaksa. Moran's I_i is statistically significant in the orange districts. In the grey districts, the darker shades correspond to lower p-values, to higher significance levels.*

Note that while the limiting distribution is Normal, the small sample distribution is not. The "sample size" is *not* the number of geographic entities being investigated; it is the number of neighbors each has. Thus, it is unlikely in real situations for the sample size to be "large enough." As such, Monte Carlo simulation should be used to estimate the p-values and the confidence intervals. Monte Carlo simulation should *also* be used to estimate the p-values and confidence intervals when the measurements are not Normally distributed, such as when they are small proportions.

To illustrate these points, let us return to the maps of Figure 4.2, which show the invalidation rate (left) and the vote proportion for Rajapaksa (right). Earlier, I

	Estimate	Std. Error	t-value	p-value
Intercept	0.0395	0.0030	13.08	< 0.0001
Candidate support	-0.0512	0.0057	-8.99	< 0.0001

Table 4.1: *The regression table for regressing the invalidation rate on the candidate support rate for the 2010 Sri Lankan presidential election. This global model is fit using ordinary least squares regression.*

remarked that there appeared to be evidence of hot and cold spots in the maps. The north seemed a hot spot for the invalidation rate and a cold spot for Rajapaksa support. Figure 4.5 shows maps of Moran's local I_i statistic for each of the districts, using rook contiguity. The left panel shows it for the invalidation rate; the right panel, the vote share for Rajapaksa. Districts colored in orange show statistically significant positive spatial correlation; districts in grey do not. Darker shades of grey indicate lower p-values (higher levels of statistical significance).

Note that our observations in Section 4.1 were correct. There is evidence of significant spatial correlation in this election in the north. Neither the invalidation rate nor the candidate support are independently distributed across the island. This result, in itself, is interesting. However, it is not a violation of the regression assumptions of Chapter 3, *per se*.

Continuing the ideas of last chapter, let us regress the invalidation rate on the candidate support rate using ordinary least squares regression. Recall from Section 3.2 that a statistically significant effect suggests unfairness in the election. The OLS-estimated effects are provided in Table 4.1. Note that the candidate support effect is highly significant. This means that one should conclude that the invalidation rate and the candidate support rate are not independent, which strongly suggests the presence of electoral fraud. However, least squares regression (OLS and WLS) assumes the residuals are independent and identically distributed. If there is spatial correlation in the residuals, this assumption is violated.

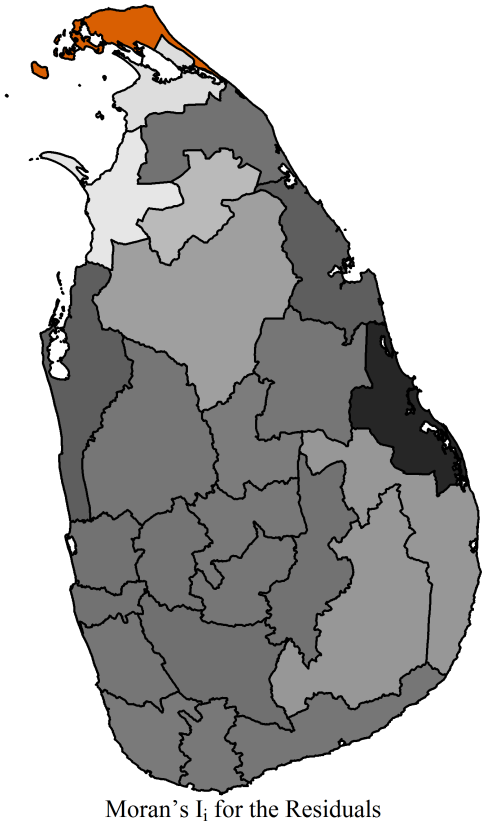


Figure 4.6: *Map of Moran's local I_i for the residuals of the global model. Moran's I_i is statistically significant in Jaffna District, colored orange. In the grey districts, the darker variations correspond to lower p-values, to higher significance levels.*

Figure 4.6 provides a map of p-values for the Moran's I_i measures. In the grey districts, the statistic is not significantly different from its expected value under the hypothesis of no spatial correlation, with darker variations indicating higher levels of significance (lower p-values). In Jaffna District (colored orange), the measure is significantly different from its expectation. This suggests that the residuals are spatially correlated. In fact, $I_9 = 5.764$, with a p-value of less than 1×10^{-5}. This indicates highly positive spatial correlation in the Jaffna District even using the Bonferroni correction. Thus, we can conclude that the residuals violate one of the least squares assumptions.

And so, we begin discussing modeling in space. The next section introduces a first attempt at reducing—and modeling—the residual spatial correlation. It is called

the spatial-lag model, and its logic follows that of its time-series regression (analysis) analogue.

4.3. The Spatial-Lag Model

In time series analysis, one of the first attempts to model the temporal correlation was the lagged dependent variable model (Wei 2006). This model included the temporally lagged dependent variable as an independent variable. The coefficient on the lagged dependent variable is termed the autocorrelation factor. Unfortunately, using ordinary least squares to estimate this factor produces biased results (Keele and Kelly 2006).

The analogue in spatial analysis is the spatial-lag model (LeSage and Pace 2009). It, too, includes a lagged dependent variable as an independent variable. The lagged dependent variable, however, is the neighborhood average. Thus, let $\mathbf{W}^*$ be the row-standardized neighborhood (adjacency) matrix; that is, define $\mathbf{W}^* := \frac{\mathbf{W}}{\mathbf{WJ}}$, where the multiplication is usual matrix multiplication, the division is Hadamard division (element-wise division), and $\mathbf{J}$ is a matrix of all 1s. The spatial-lag model is

$$\mathbf{Y} = \rho\mathbf{W}^*\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Here, ρ is the neighbor effect, also known as the contagion effect.

To see that the OLS estimates of ρ can be biased, let us simplify this model significantly. That is, let us define $\bar{y}_{(i)}$ as the arithmetic mean of the neighbors of entity i. Let ρ be the level of spatial correlation between neighboring entities. The simplified spatial-lag model, with $\rho \in (-1, 1)$, is

$$y_i = \rho\bar{y}_{(i)} + \varepsilon_i \tag{4.2}$$

With this simplification, this proof follows Anselin (1988), showing the bias.

Theorem 4.3 (Bias in the Spatial-Lag Model)**.** *Let us be given Model 4.2, with the errors* $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$*. If the residuals are spatially correlated, the OLS estimate for* ρ *is biased.*

Proof. As the model is $y_i = \rho\bar{y}_{(i)} + \varepsilon_i$, the OLS estimator for ρ is

$$\hat{\rho} = \frac{\sum_{i=1}^{n} y_i \bar{y}_{(i)}}{\sum_{i=1}^{n} \bar{y}_{(i)}^2}$$

If we substitute (4.2) into the numerator and expand, we have

$$\begin{aligned}
\hat{\rho} &= \frac{\sum_{i=1}^{n} \rho\bar{y}_{(i)}^2 + \sum_{i=1}^{n} \varepsilon_i \bar{y}_{(i)}}{\sum_{i=1}^{n} \bar{y}_{(i)}^2} \\
&= \frac{\rho \sum_{i=1}^{n} \bar{y}_{(i)}^2}{\sum_{i=1}^{n} \bar{y}_{(i)}^2} + \frac{\sum_{i=1}^{n} \varepsilon_i \bar{y}_{(i)}}{\sum_{i=1}^{n} \bar{y}_{(i)}^2} \\
&= \rho + \frac{\sum_{i=1}^{n} \varepsilon_i \bar{y}_{(i)}}{\sum_{i=1}^{n} \bar{y}_{(i)}^2}
\end{aligned}$$

Thus, $\mathbb{E}[\hat{\rho}] = \rho$ if $\mathbb{E}\left[\varepsilon\bar{y}_{(\cdot)}\right] = 0$; that is, the estimator is unbiased if the errors are uncorrelated with the neighborhood averages. In reality, this strict exogeneity is never fully met. Thus, if the entity measurements are spatially correlated, then so are the neighborhood averages. As we assumed the errors ε_i are independent and identically distributed, $\mathbb{E}\left[\varepsilon\bar{y}_{(\cdot)}\right] = 0$ only when the residuals are uncorrelated. □

Thus, the parameter estimates from the spatial lag model are biased. Theorem 4.3, however, does not specify the level of bias. In time series analysis, (Keele and Kelly 2006, p186) showed that the lagged variable model was biased, but not much:

> while the lagged dependent variable is inappropriate in some circumstances, it remains the an appropriate model for the dynamic theories often tested by applied analysts

In light of their analysis, it appears as though the spatial-lag model is an appropriate model for this research. Its parameter estimates are only slightly biased, and it does reduce the level of spatial correlation. Finally, it can be used with any of the regression techniques covered in the previous chapter.

4.3.1 The SLM and Sri Lanka. To demonstrate the spatial-lag model, let us use it on the Sri Lankan 2010 presidential election. Recall that the dependent variable is the

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.0257	0.0055	4.69	0.0001
Neighbor effect	0.4384	0.1535	2.86	0.0092
Candidate support	-0.0354	0.0074	-4.76	0.0001

Table 4.2: *Regression table for the spatial-lag model. Note that the neighbor effect is statistically significant, as is the candidate effect.*

invalidation rate in each division and the independent variable is the candidate support rate for incumbent Rajapaksa in the division.

The choice of neighborhood is uncomplicated as there are no bishop-contiguous districts. Thus, using rook neighborhoods is equivalent to using queen neighborhoods. Furthermore, LeSage and Pace (2012) concluded that the parameter estimates are *not* sensitive to the choice of neighborhoods because the neighborhoods themselves are highly correlated.

The parameter estimates from fitting the spatial-lag model are presented in Table 4.2. Note the statistical significance of the neighbor effect, $\hat{\rho} = 0.4384$. This finding suggests that the neighborhood affects the invalidation rate. This result is consistent with the hypothesis that geography matters.

There is no evidence of spatial correlation in the residuals of this model. The lowest p-value is 0.2370, which is for Jaffna District in the extreme north of the island. Thus, *this* assumption of ordinary least squares regression is not violated. Whether the assumption that the coefficients are constant is violated cannot be tested using this model.

In terms of the free and fair hypothesis, this model provides evidence against it. Here, the invalidation rate and the candidate support rate are not independent ($p = 0.000095$). The effect, however, is rather small. For every 1% increase in the vote proportion for Rajapaksa, the invalidation rate drops by 0.035%. For a 20% increase in Rajapaksa support, the invalidation rate drops by 0.71%.

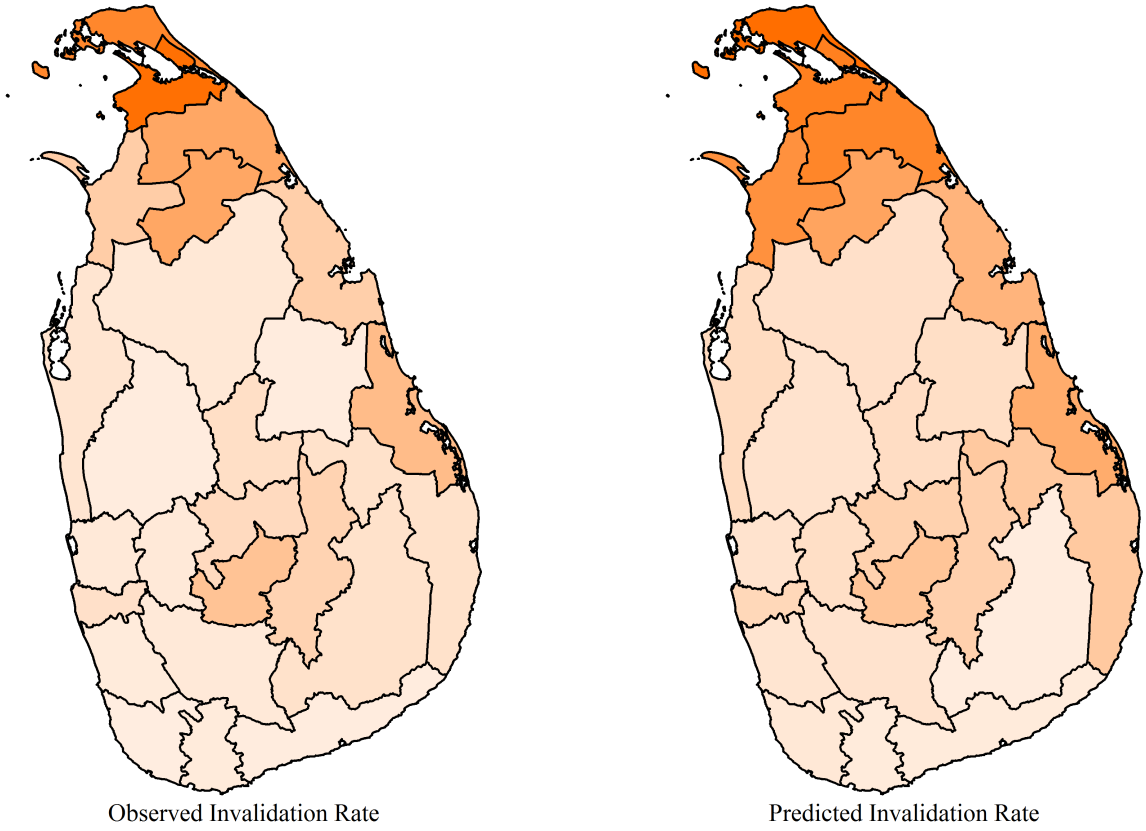


Figure 4.7: *Maps of the invalidation rate. The left map is the reported invalidation rate. The right map is the predicted invalidation rate using the spatial-lag model.*

Figure 4.7 provides maps of the reported invalidation rate (left) and the predicted invalidation rate (right). Note that the model did well throughout the island, except for the northern districts. One of the drawbacks to the spatial-lag model is that it is based on neighbor averages. The Jaffna District has only one neighbor, Kilinochchi District. All other districts have at least three neighbors.

4.3.2 Conclusion. The spatial lag model is effective at reducing the spatial correlation, which is an improvement over the ordinary least squares regression of the previous chapter. However, it has one important drawback. The spatial lag model makes the assumption that the parameter effect is constant across the entire map, that the effect surface is horizontal. This is a strong assumption that must be tested.

Thus far, we have not come across a test that allows for spatially varying effects. To allow for non-constant effects, Casetti (1972) replaced the constants with functions of location. This I explore in the next section.

4.4. Casetti's Spatial Expansion Method

In the previous section, I introduced the spatial-lag model, which is used for two purposes. First, it can reduce the spatial correlation in the residuals. Second, it models the neighborhood effect. It is not alone in the first of these two goals. Casetti (1972) created the spatial expansion method to also reduce the spatial correlation of the residuals. More importantly, the expansion method also models spatially-dependent *effects*—effects varying across the map.

The spatial-lag model for a single predictor variable is

$$y_i = \rho \bar{y}_{(i)} + \beta_0 + \beta_1 x_i + \varepsilon_i$$

The model assumes that β_1 effect is constant throughout the entire region; it is spatially invariant. To model spatially *varying* effects, we replace the constant β_1 with an arbitrary function $\beta_1(i)$, where i is the entity and is a function of location (u, v); that is, $\beta_1(u, v)$ describes the effect surface. From this, the expansion model is

$$y_i = \beta_0(u, v) + \beta_1(u, v)x_i + \varepsilon_i \tag{4.3}$$

The complexity of the functions is limited only by the number of parameters needed to estimate and the sample size—the degrees of freedom. This model is fit using any of the regression techniques of Chapter 3. The issue reduces to determining the correct functions. This is the method's weakness (Fotheringham et al. 1998). Rarely is there *a priori* knowledge about how the effects vary.

Using quadratic functions allows 4.3 to model effects that have a single extremum on the map. Cubic functions can model effects with two extrema. Higher order functions

can model effects with more extrema, at the expense of degrees of freedom and of fitting the data rather than the process (the bias-variance tradeoff). As the purpose of this research is to detect any spatial variation in the effects, a quadratic function may be sufficient.

4.4.1 The SEM and Sri Lanka. To illustrate this method, let us return to the Sri Lanka election and, without prejudice, use quadratic functions for $\beta_0(u, v)$ and $\beta_1(u, v)$. The model is

$$y_i = \Big(a_0 + a_1 u + a_2 v + a_3 uv + a_4 u^2 + a_5 v^2\Big) + \Big(b_0 + b_1 u + b_2 v + b_3 uv + b_4 u^2 + b_5 v^2\Big) x_i + \varepsilon_i \tag{4.4}$$

Here, u and v are arbitrary positional variables. I will use longitude and latitude for the position of the division centroid. Using quadratic functions ensures that at most a single internal extremum will be detected. In general, election theory should be used to select the correct polynomial degree. However, such theory rarely exists.

Using ordinary least squares regression, the estimated effects, standard errors, t-values, and p-values are provided in Table 4.3. Note that the nominal invalidation rate is not constant across the island. From Figure 4.8, we see it has a minimum of 0.0057 in Puttalam District in the far west and a maximum of 0.0698 in Hambantota District in the south.

The candidate effect also spatially varies. It is strongest in the south (Puttalam, -0.0936) and weakest in the middle west (Hambantota, 0.0029). It is also strong in the north of the island, where the candidate effect is -0.0656 in Jaffna District. This direction of variation is evident from the regression results (Table 4.3). The north-south quadratic effect (v^2) is statistically significant and has a higher magnitude than that of the east-west effect (u and u^2). It may also be interesting to note that the north-south quadratic effect is the *only* statistically significant effect.

	Estimate	Std. Error	t-value	p-value
Nominal Effect				
Intercept	-64.6802	107.8748	-0.60	0.5591
u	1.5834	2.5862	0.61	0.5509
v	0.0778	1.0478	0.07	0.9419
uv	-0.0036	0.0123	-0.30	0.7724
u^2	-0.0096	0.0155	-0.62	0.5488
v^2	0.0135	0.0050	2.70	0.0182
Candidate Effect				
Intercept	54.1632	198.5341	0.27	0.7893
u	-1.3505	4.7840	-0.28	0.7822
v	0.3137	1.7839	0.18	0.8631
uv	0.0003	0.0211	0.01	0.9898
u^2	0.0082	0.0289	0.28	0.7805
v^2	-0.0211	0.0082	-2.57	0.0231

Table 4.3: *Results of the expansion method regression as described in the text. Note that there is significant positional effect on the invalidation rate (top block) and on the invalidation rate through the candidate support rate (bottom block).*

More importantly, this OLS model also shows no significant spatial correlation in the residuals. Thus, there is no apparent violation of the assumption of the spatial independence of the residuals. The expansion model allowed us to model the varying effects of the candidate support *and* eliminated the troubling spatial correlation in the residuals.

There is no *a priori* reason to select quadratic functions for the positional effects. The fact that the north-south quadratic effect is the only statistically significant effect is due solely to the distribution of the candidate effect on the ground.

4.4.2 Conclusion. The major drawback of this method is that the functional form of the effect parameters is rather constrained. A high-degree polynomial will allow the estimated effect function to better match the true variation. However, a high-degree polynomial will also tend to reduce the degrees of freedom too much. Furthermore, in both cases, there is the issue of model misspecification. The p-values are only appropriate

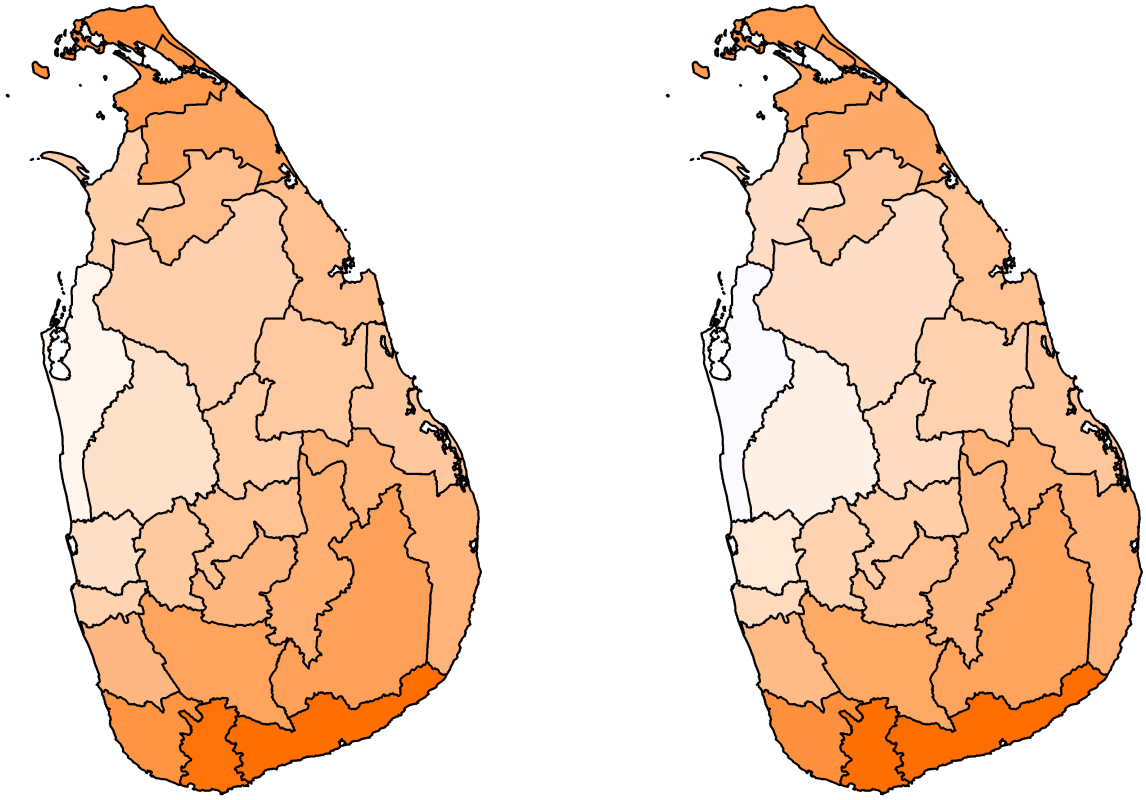


Figure 4.8: *Map of spatially-varying effects. The left map is of the base invalidation rate. The right map is of the vote proportion for Rajapaksa across Sri Lanka. Darker tones of orange correspond to high negative effects. White corresponds to no effect. Darker tones of blue correspond to high positive effects. The district with the most extreme candidate effect was Matara District in the south of Sri Lanka, with* $\beta_1 = -0.0936$.

if the model is correctly specified. To achieve the goal of flexibility in the effect function without the cost in terms of degrees of freedom, Fotheringham created the geographically weighted regression model.

4.5. Geographically Weighted Regression

A. Stewart Fotheringham, Martin E. Charlton, and Chris Brunsdon (1998) saw the expansion method as too rigid. They sought to allow the data a larger role in determining the shape of the effect surface. This would allow for an arbitrary number of local extrema in the effect surface.

Name	Formula
Unweighted	$w_{ij} = 1$
Local unweighted	$w_{ij} = \mathbb{1}\{d_{ij} \leq h\}$
Gaussian	$w_{ij} = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)\right]$
Bisquare	$w_{ij} = \left(1 - \left(\frac{d_{ij}}{h}\right)^2\right)^2 \mathbb{1}\{d_{ij} \leq h\}$
Tricube	$w_{ij} = \left(1 - \left(\frac{d_{ij}}{h}\right)^3\right)^3 \mathbb{1}\{d_{ij} \leq h\}$
Epanechnikov	$w_{ij} = 1 - \left(\frac{d_{ij}}{h}\right)^2 \mathbb{1}\{d_{ij} \leq h\}$

Table 4.4: *List of some kernels commonly used in geographically weighted regression. In all formulas, d_{ij} is the distance between point i and point j, and h is the bandwidth for the kernel.*

Geographically weighted regression estimates the effects at each point on the map using weighted regression. These points may be points representing the entities, or those points may be elements of an overlaying grid. The weighting scheme is selected by the researcher, but should conform to the rules of kernel estimations. As with kernel density estimation, the choice of kernel function is largely arbitrary, but the choice of bandwidth is important (Fotheringham et al. 2003, Chapter 2). Table 4.4 provides several common kernels, along with the corresponding functions. To calculate the weights matrix at location i, you calculate the w_{ij} for your selected kernel and bandwidth, $\mathbf{W}_i = diag\{w_{ij}\}$ for j ranging across all positions.

In terms of regression, geographically weighted regression looks familiar. The model is

$$\mathbf{Y} = (\boldsymbol{\beta}' \circ \mathbf{X})\,\mathbf{j} + \boldsymbol{\varepsilon}$$

Here, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors, $\mathbf{j}$ is an $(p+1) \times 1$ vector of 1s, $\circ$ is the Hadamard product (element-wise multiplication), $\mathbf{X}$ is the $n \times (p+1)$ design matrix, and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0(1) & \beta_0(2) & \cdots & \beta_0(n) \\ \beta_1(1) & \beta_1(2) & \cdots & \beta_1(n) \\ \vdots & \vdots & \ddots & \vdots \\ \beta_p(1) & \beta_p(2) & \cdots & \beta_p(n) \end{bmatrix}_{(p+1)\times n}$$

Recall that n is the number of locations involved in the analysis.

Without additional structure, this equation cannot be solved; the number of parameters to estimate, $n \times (p+1)$, is greater than the number of data values to use. The weights matrix adds this structure.

If we focus on a single location, i, the model equation becomes the familiar $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}(i) + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}(i)$ is the ith column of $\boldsymbol{\beta}$. At this point, Fotheringham et al. (2003, Chapter 2) states

$$\hat{\boldsymbol{\beta}}(i) = \left(\mathbf{X}'\mathbf{W}_i\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}_i\mathbf{Y}$$

where $\mathbf{W}_i$ is a diagonal matrix consisting of the ith column of the neighbor matrix. However, this result is not always correct. The correct result is the more general

$$\hat{\boldsymbol{\beta}}_i = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| \mathbf{W}_i^{1/2}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right) \right\|^2,$$

which comes from the definition of least squares.

The Fotheringham et al. (2003) result is appropriate only when the weights are the inverse of the variances associated with each measurement. In general, the $\mathbf{W}_i$ has nothing to do with the variances. The theorem of Section 3.2.2 on page 66 shows the condition under which the Fotheringham et al. (2003) result is valid.

The Fotheringham et al. (2003) result also requires the neighbor matrix to be positive definite. This requirement is rarely an issue, unless you are investigating maps with islands. Using any contiguity measure on the 50 US states results in a neighbor matrix with rank less than 50; neither Hawai'i nor Alaska has a neighbor. To solve this issue, cell values tend to be functions of distances rather than of contiguity.

Note the most important feature of geographically weighted regression: the parameter estimates are allowed to vary at each experimental unit. This allows one to explore how the effect varies across the map. Unfortunately, even when using the Fotheringham et al. (2003) result, there is no native way of conducting hypothesis tests or of creating confidence intervals; the number of degrees of freedom are not known.

4.5.1 The GWR and Sri Lanka. To illustrate geographically weighted regression, let us once again return to the 2010 presidential election in Sri Lanka. As the kernel function is of little importance, I arbitrarily select the Gaussian kernel. In Kernel density estimation, the bandwidth is much more important (Epanechnikov 1969). LeSage and Pace (2012), however, imply that the final regression results are robust to even the bandwidth.

For this example, I selected the fixed bandwidth that produced the lowest Akaike's Information Criterion (Akaike 1974). This resulted in a fixed bandwidth of 150km. This means that each weighted regression includes approximately five neighbors.

According to the global model fit previously, the candidate effect is a constant -0.0512 (Table 4.1). According to the spatial expansion model, the candidate effect varies from -0.0940 to $+0.0029$. According to this geographically weighted regression model, the candidate effect varies from -0.05699 (Jaffna District in the extreme north) to -0.03904 (Hambantota District in the south).

Note that this result substantively differs from that of the spatial expansion model (Section 4.4.1). That model concluded that the candidate effect was lowest in the center of the country, with maxima along the north and south coasts. Here, the maximum is still along the north coast. The minimum, however, is now estimated to be the southern coast. Figure 4.2 may shed some light on the two models. The spatial expansion method fit assumed a quadratic effect. The observed invalidation rate (Figure 4.2) is high in the north and east, but it varies little across the rest of the districts. The quadratic function

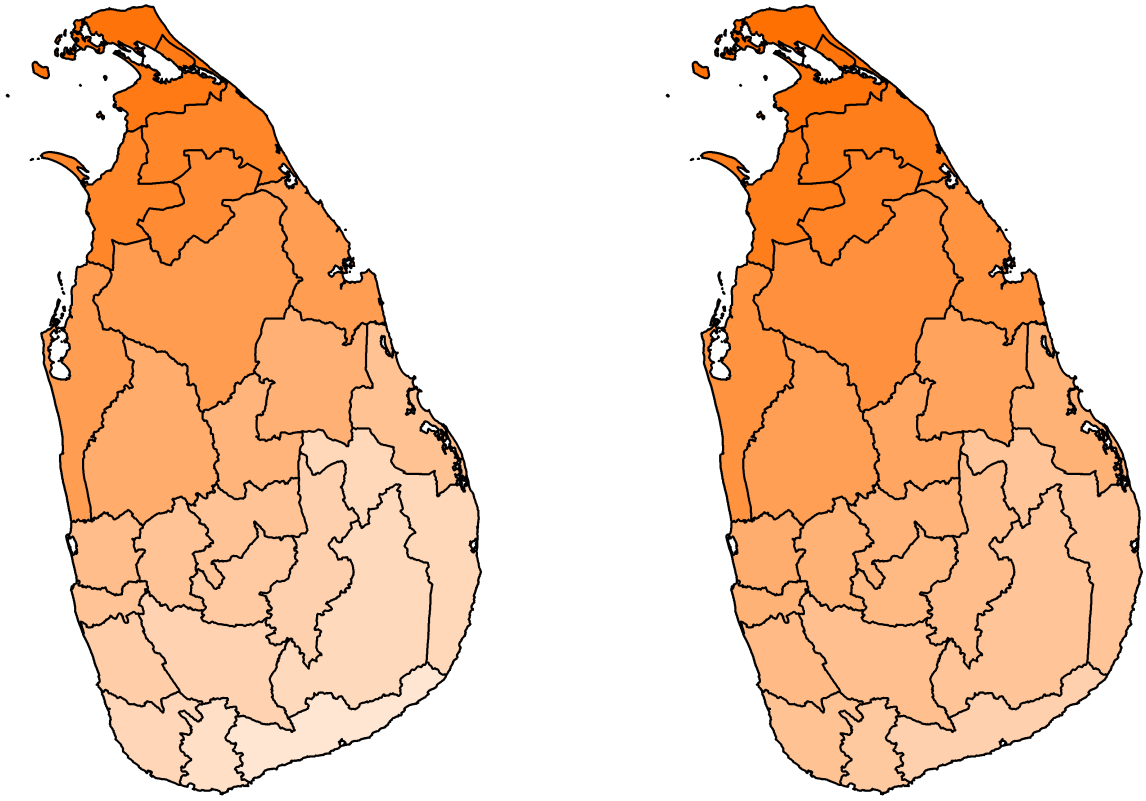


Figure 4.9: *Map of spatially-varying effects for the geographically weighted regression model. The left map is of the base invalidation rate. The right map is of the vote proportion for Rajapaksa across Sri Lanka. Darker tones of orange correspond to high negative effects. White corresponds to no effect. The candidate effects range from* -0.0570 *in Jaffna District to* -0.0390 *in Hambantota District.*

forced the specific shape in the effect surface. Geographically weighted regression does not force a specific shape; it allows the observations to create the shape.

Figure 4.9 shows the spatially varying effects. The left panel shows the variation in the nominal invalidation rate; the right panel, the variation in the candidate effect. Both maps suggest the process generating the invalidation rate in the north is dissimilar to that in the south. Unfortunately, there is no native method for determining if that difference is statistically significant or if the effect is statistically significant. The weighting invalidates the usual rules for degrees of freedom. As such, bootstrapping remains the feasible method for estimating confidence intervals and p-values (Fotheringham et al. 2003).

4.5.2 Conclusion. In geographically weighted regression, weighted regression is performed at each point i. This means the estimated effects parameter $\hat{\boldsymbol{\beta}}$ is a function of the point i; that is, they are spatially varying. Furthermore, as points are weighted and used in separate regressions, it means the number of degrees of freedom are *not* the sample size n less the number of parameters fit.

The first result was the goal of Fotheringham et al. (1998). The effect surface can now be more closely estimated. The second result is the most serious drawback of geographically weighted regression. Until the number of degrees of freedom can be determined, unadjusted t-tests should not be performed. Fotheringham et al. (2003, §2.8) states this.

Regardless of the lack of tests and of the warning of Fotheringham et al. (2003), researchers use geographically weighted regression to test hypotheses. Chen and Truong (2012) attempt to find a relationship between obesity and township disadvantages in Taiwan. They use geographically weighted regression, testing for parameter significance and parameter variation using critical values of 1.96 and 3.92. Li et al. (2010) examine the relationship between temperatures various environmental factors. The authors used the typical t-tests to determine the statistical significance of their various variables.

Regardless of the technique's (mis)use, geographically weighted regression does compare favorably to the expansion method in one important manner (Paéz 2005). Geographically weighted regression is more flexible in estimating the effect surface than is the expansion method. That this was the purpose behind geographically weighted regression is a positive. However, many researchers do warn about the method's weaknesses (Bivand and Yu 2013; Fotheringham et al. 2003; Paéz 2005).

In the next section, I create a method that retains the ability to test hypothesis, while increasing the flexibility of the spatial expansion method. I term it the lagged expansion method.

Fotheringham et al. created geographically weighted regression to allow the data to strongly suggest the effect surface. It was *intended* more for exploratory analysis than for inferential analysis (Fotheringham et al. 1997, 2003). This idea is echoed by Bivand and Yu (2013), in R's `spgwr` package, which warns

```
NOTE: This package does not constitute approval of GWR
as a method of spatial analysis
```

As seen in the previous section, geographically weighted regression provides much by way of pretty maps, but little by way of testing. Casetti's expansion method provides the testing methods, but is limited in its flexibility. The spatial lag model is flexible in allowing local variation, but it does not allow for spatially varying effects. My solution is to combine two of these three to create a method superior to the third. Adding the spatial lag to the expansion method will allow flexibility approaching that of geographically weighted regression, while still allowing for hypothesis testing, which is required in this research.

And so, the spatially-lagged extension model is

$$y_i = \rho(u,v)\bar{y}_{(i)} + \beta_0(u,v) + \beta_1(u,v)x_i + \varepsilon_i \tag{4.5}$$

Here, $\rho(u,v)$ is the neighbor (contagion) effect, $\bar{y}_{(i)}$ is the neighbor-average for entity i, $\beta_0(u,v)$ and $\beta_1(u,v)$ are the nominal effect and the candidate effect, and ε_i is the error, $\varepsilon \sim \mathcal{N}_n(\mathbf{0}, \mathbf{\Sigma})$.

Note that this model adds several features. First, it incorporates both spatially-varying effects and a neighbor effect. Second, it allows for the neighbor effect to vary across the map. Third, it no longer requires the errors to be independent and identically distributed.

Fitting this model is easy in light of the regression methods of Chapter 3, specifically feasible generalized least squares regression of Section 3.2.3.

	Estimate	Std. Error	t-value	p-value
Neighbor Effect				
Intercept	-1219.2381	665.0389	-1.83	0.1000
u	15.2309	8.2733	1.84	0.0988
v	161.9476	72.7705	2.23	0.0531
uv	-2.0235	0.9062	-2.23	0.0524
Nominal Effect				
Intercept	-104.2957	93.6622	-1.11	0.2943
u	3.0057	2.2477	1.34	0.2140
v	-4.6213	1.5891	-2.91	0.0174
u^2	-0.0210	0.0135	-1.55	0.1545
v^2	0.0232	0.0067	3.48	0.0070
uv	0.0531	0.0191	2.78	0.0214
Candidate Effect				
Intercept	171.3883	195.4682	0.88	0.4034
u	-4.6963	4.7776	-0.98	0.3513
v	4.9102	1.8895	2.60	0.0288
u^2	0.0316	0.0292	1.08	0.3078
v^2	-0.0308	0.0084	-3.68	0.0051
uv	-0.0550	0.0225	-2.44	0.0371

Table 4.5: *Regression table for the results of the spatial lagged expansion method. As with the spatial expansion method, u represents the longitude; v, the latitude.*

4.6.1 The SLEM and Sri Lanka. Returning again to the 2010 Presidential election in Sri Lanka, I fit a model with all effects at the quadratic level. However, the statistical significance of the parameter estimates indicated the model was overfit. Because of this, I used a linear neighbor effect, a quadratic nominal effect, and a quadratic candidate effect for the model. The results of the regression are provided in Table 4.5.

Note that there remain quadratic effects for both the nominal effect and for the candidate effect. Note also that this effect is more pronounced in the north-south direction (v) than in the east-west (u). The adjusted R^2 for this model is 0.95, and its AIC is -228. The adjusted R^2 for the global model is only 0.77, with an AIC of -192. The adjusted R^2 for the expansion model is only 0.82, with an AIC of -198. Thus, by these measures, the spatial lag expansion method is superior.

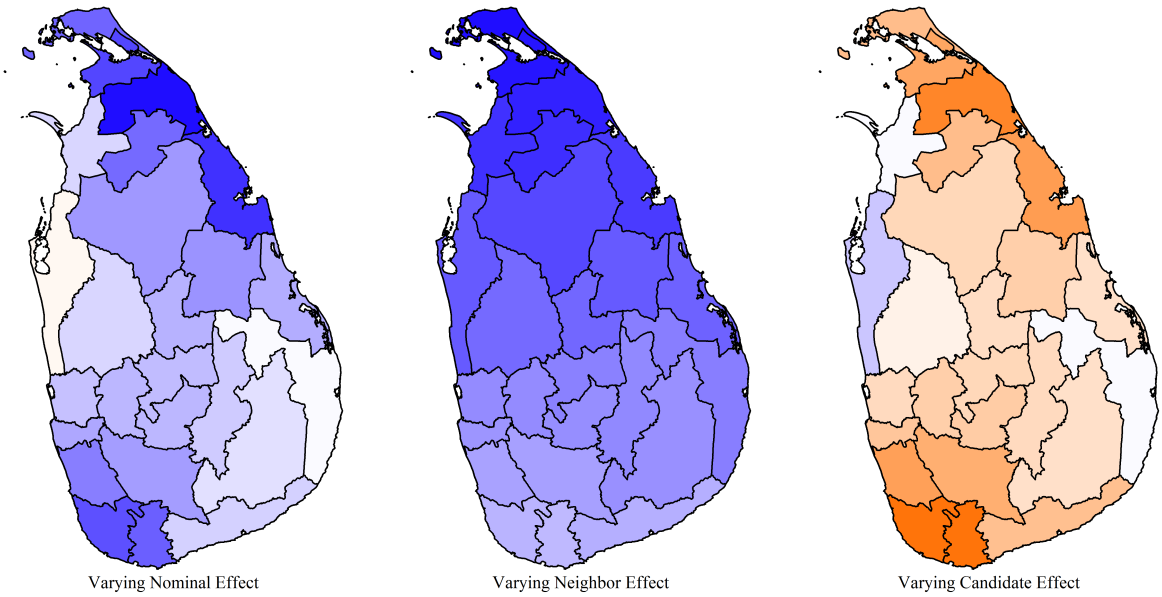


Figure 4.10: *Map of spatially-varying effects for the spatially lagged expansion model. The left map is of the base invalidation rate. The center map is of the neighbor effect. The right map is of the vote proportion for Rajapaksa across Sri Lanka. Darker tones of orange correspond to high negative effects, white to no effect, and darker tones of blue to positive effects.*

Comparing these results to that of the geographically weighted regression model tells a similar story. For the geographically weighted regression model, the (approximate) adjusted R^2 is 0.61; the AIC is -199; and the local R^2 values range from 0.75 to 0.80.

To be useful, the spatial lagged expansion method also needs to be able to mitigate any spatial correlation in the residuals. In this case, it did so. The lowest p-value for the Moran's local I_i measure is 0.2438. Thus, where the global model had spatial correlation in the residuals, this model does not.

Figure 4.10 shows maps of the varying effects across Sri Lanka. The right map is of the candidate effect. Note that the candidate effect is strongest in both the north and the south, and both are negative. This negative effect strongly implies a significant relationship between the invalidation rate and the candidate support rate in those areas. This is a violation of the free and fair hypothesis. That the values are negative indicates higher candidate support corresponds to a lower invalidation rate. This is consistent with the effects of ballot box stuffing and some other forms of election fraud.

The left map is of the nominal invalidation rate, with shades of blue signifying a positive effect. Note that there are high nominal invalidation rates in both the north and the south. This map implies that the base invalidation rate is high in the north and the south. This is a also a violation of the free and fair hypothesis. These results are consistent with an inherent unfairness in the electoral system. Future analysis should focus on how those two regions are different from the middle region.

Finally, the middle map is of the neighbor effect, also known as the contagion effect. Note that this is always positive and is strongest in the north. While this is not directly a violation of the free and fair hypothesis, it does raise a troubling question: Why is the data-generating process in the north different from that in the south?

4.6.2 Conclusion. In this section, I introduced the spatial lag expansion model—a geographically-based method that combines the strengths of the spatial lag model and the expansion method while keeping the native ability to perform hypothesis tests, which the geographically weighted regression model does not have.

With the new method, I again tackled the 2010 Sri Lankan presidential election to test the official results for violations of the free and fair hypothesis. Again, we find the election falls short of the democratic goal. The fact that this model allows for a spatially varying contagion effect is a strength.

4.7. Type I and Type II Error Rates

While the spatial lag expansion method did seem to exceed the performance of the current three methods, this improvement may simply be a matter of the Sri Lankan data. As with previous chapters, the methods must be tested against two benchmarks: the Type I Error rate and the power.

In previous chapters, testing the Type I Error rate was relatively easy. The null hypothesis is the independence of the invalidation rate and the candidate support. Here,

there is the added issue of spatial correlation. The null hypothesis does not concern this. As such, the Type I Error rate needs to be tested using several levels of spatial correlation. As its effect on estimation appears to be minor, I dismiss it below.

Also new to this chapter is the importance of positional relationships. To investigate how this affect power and other aspects, I use three maps: Belgium (GADM 2014a), Sweden (GADM 2014a), and a regular grid (Anselin et al. 2005). I selected Belgium as the number of electoral divisions is small (11 provinces), forcing me to examine small-sample properties. I selected Sweden because the divisions (21 counties; *län*) had few neighbors in the north. I selected the 5×5 regular grid as a counterpoint to Sweden in terms of connectivity and to Belgium in terms of size.

4.7.1 Effects of Spatial Correlation. Varying levels of spatial correlation can be generated using a bivariate Normal distribution. That is, if the positions of the n electoral divisions are $(\mathbf{x}, \mathbf{y})_i$, then one can generate spatially correlated data using

$$\mathbf{C} \sim \mathcal{N}_n \left(\boldsymbol{\mu}; \sigma^2 \boldsymbol{\Sigma} \right)$$

where $\mathbf{C}$ is the variable being generated (the candidate support rate), $\boldsymbol{\mu}$ is a vector of expected values, σ^2 is the nominal variance, and $\boldsymbol{\Sigma}_{ij} = \exp\left[-\phi d_{ij}\right]$. Here, ϕ is a measure of spatial correlation and d_{ij} is the distance between electoral divisions i and j. If $\phi \leq 0$, then the correlation matrix is singular.

Once the spatial correlation level is specified, independently-generated invalidation rates and candidate support rates can be generated, each generation being a new simulated election. This I do for correlation values of $\rho = 0.0$, 0.5, and 0.9 (Figure 4.11). It is important to note that the power curves do not appear to vary due to changing correlation.

While this graphic is only for varying the contagion effect for the 5×5 grid, the effects were similarly interesting for the maps of Belgium and Sweden. This is good as

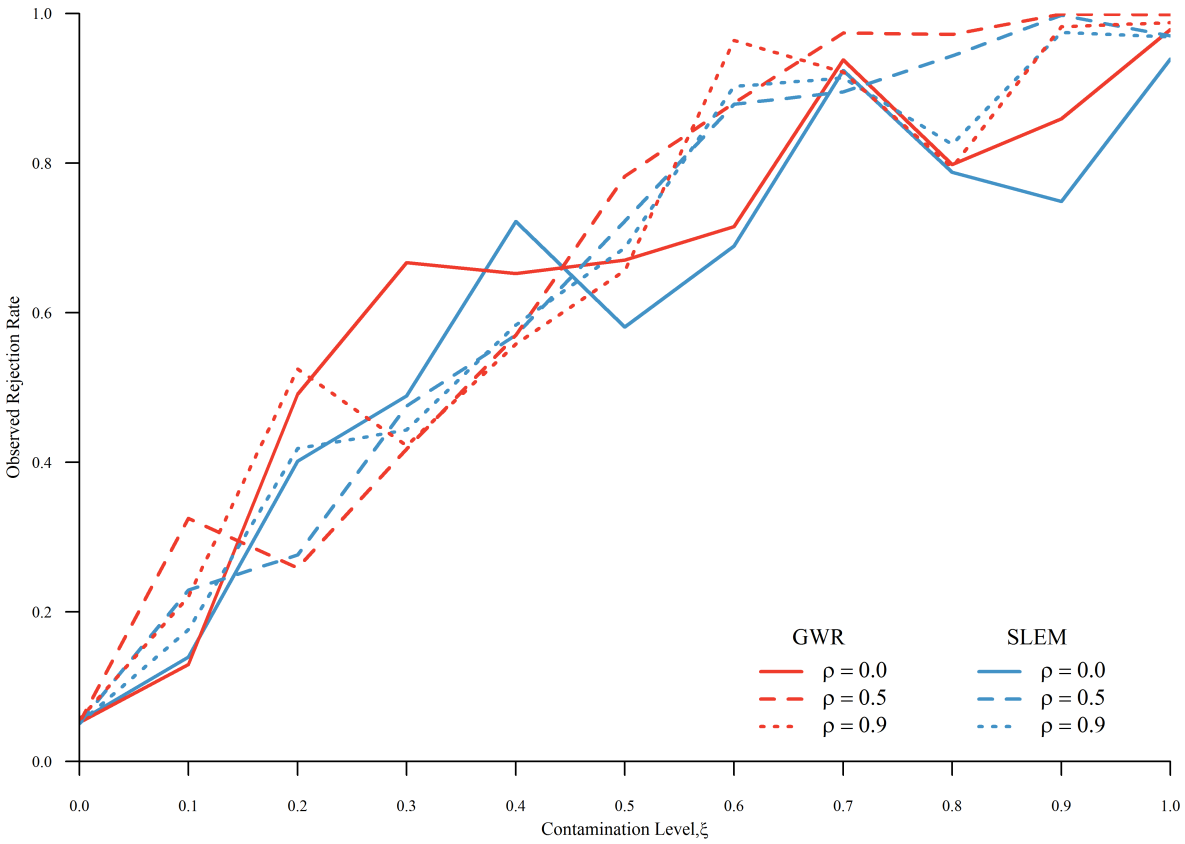


Figure 4.11: *Power curve for the effect of contamination on the rejection rate, separated by estimation method and level of contagion. Note that neither estimation method* nor *contagion level affects the power curve.*

it means these two estimation methods are robust to the contagion effect. It also means one fewer variable to investigate in power calculations.

4.7.2 The Test Statistics. While I suggested that the spatially lagged expansion method has a natural inferential test, such is not strictly the case. Least squares regression assumes that the model is properly specified. This is problematic due to the paucity of information regarding election distributions. In Section 4.6, I suggested using a quadratic model. This is only appropriate if the quadratic model is correct. As we have no *a priori* information about what model would be correct, we cannot blindly rely on the usual t-tests, especially if we want to optimize the power of the test. In lieu of

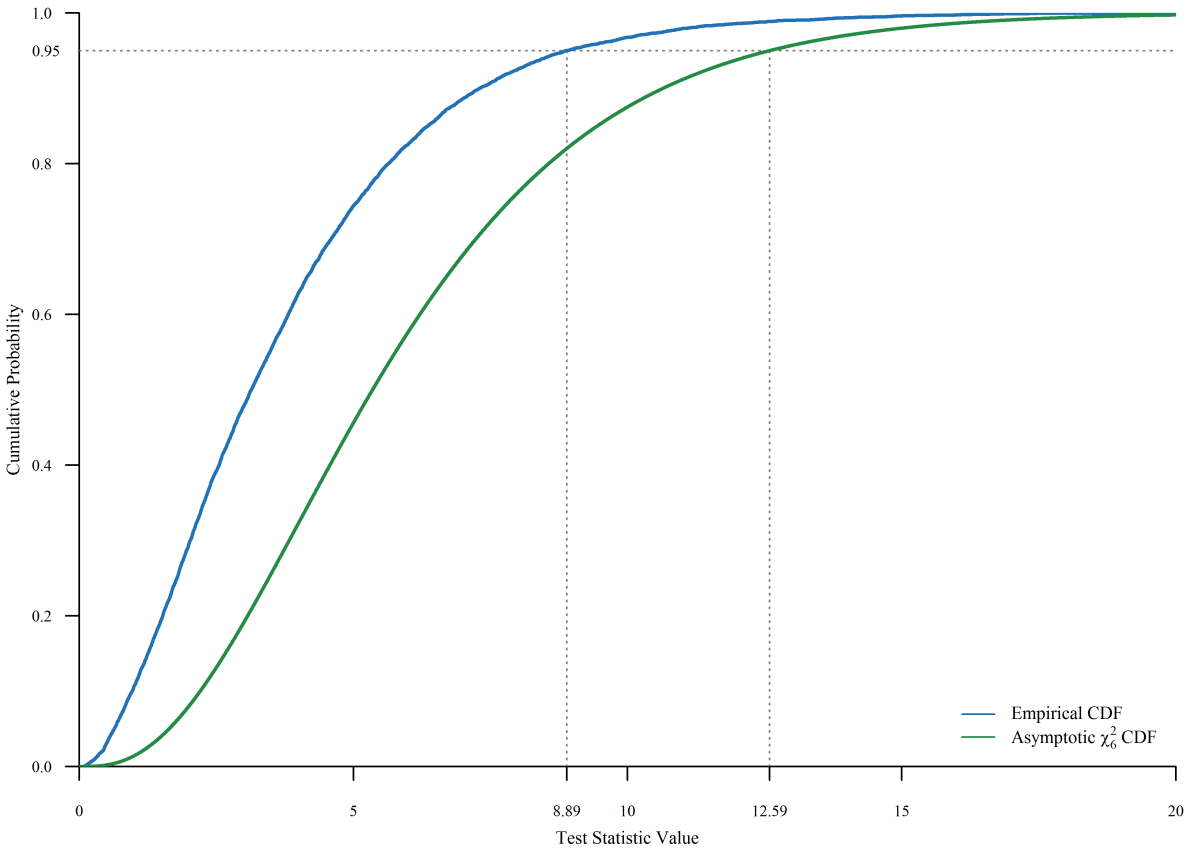


Figure 4.12: *The empirical CDF of test statistic as compared to the CDF of the χ^2_6 distribution. Note that the asymptotic distribution is stochastically greater than the observed distribution; it is of less power.*

using the asymptotic distribution, I suggest using simulation to obtain the distribution of an appropriate test statistic. In this research, I used the difference of the log-likelihood of the full model and the log-likelihood of the reduced model (the model without the candidate effect). If this difference is high, then there is support for concluding that the candidate support rate is an important predictor of the invalidation rate. If this ratio is close to 1, then there is little support for such a conclusion.

Figure 4.12 is a graphic of the empirical cumulative distribution functions for the observed test statistic and for its asymptotic distribution, χ^2_6. Note that the asymptotic distribution is stochastically less than the observed distribution. As such, it will be a conservative test of the null hypothesis. This makes it a valid, albeit weak, test. For

Country	SLEM	GWR	n
Belgium	1.5122	0.0175	11
Sweden	1.1740	0.0073	21
Grid	1.1199	0.0048	25

Table 4.6: *Estimated critical values for the test statistics of each estimation method and for each map. SLEM's is based on the ratio of the log-likelihoods of the full model to the restricted model; GWR's on the pseudo p-values. All correspond to* $\alpha = 0.05$*. The final column,* n*, provides the number of electoral divisions in the map. All critical values are estimated based on 100,000 iterations.*

the grid, one would reject when the test statistic is greater than 12.59 when using the asymptotic distribution, but at only 8.89 when using the simulated distribution.

Geographically weighted regression has similar issues. Because of the lack of true number of degrees of freedom, and due to multiple testing issues, the distribution of a test statistic must also be estimated for GWR. This test statistic needs to take into consideration the estimated degrees of freedom, the local standard errors, and the local effect estimates. To combine these three, I use the calculated pseudo p-value for each division. The final test statistic is the minimum of these pseudo p-values.

Table 4.6 provides the estimated critical values to four decimal places. These are estimated using 100,000 elections generated under the null hypothesis with zero contagion effect. Note that for GWR, the Bonferroni adjustment is very conservative, as expected.

4.7.3 Power and Irregularities. For each of the three maps, I created elections where the invalidation rate depended on two things to varying degrees: position (latitude and longitude) and candidate support. Because of the small number of divisions, I included linear functions of these, transforming to ensure that the invalidation rate was bounded between 0 and 1. Finally, as it was the independent variable, the candidate support rate was drawn from a multivariate Normal distribution with a random correlation structure, which was kept constant for the entire simulation run. Keeping the structure constant allowed me to show that the power curves are dependent on that structure.

For Sweden and the grid, the reduced model consisted of interaction terms of the latitude, longitude, and neighborhood-average invalidation rate. For Belgium, due to the number of parameters and divisions, only the additive terms were included. In all cases, the full model included these terms along with their products with the candidate support rate.

I varied the candidate effect from 0.0 to 1.0 in steps of 0.1. I gave the latitude and the longitude effects values of 0.000, 0.003, and 0.006, and their interaction effect values of 0.000, -0.003, and -0.006. For reasons discussed in Section 4.7.1, I set the contagion effect to be zero.

In summary, if u corresponds to latitude, v to longitude, ρ to contagion, and c to the candidate, I varied the parameters as

$$\begin{aligned}
\beta_c &= 0.0(0.1)1.0 \\
\beta_u &= 0.000, \quad 0.003, \quad 0.006 \\
\beta_v &= 0.000, \quad 0.003, \quad 0.006 \\
\beta_{uv} &= 0.000, -0.003, -0.006 \\
\beta_\rho &= 0
\end{aligned}$$

Note that the latitude ranges from 0.5 to 4.5 for the grid, but from 49 to 51 for Belgium and from 55 to 69 for Sweden—an increase of an order of 10. Because of this, I only used 0.0000 and 0.0001 for Belgium and Sweden for parameters β_v and -β_{uv}. That is, since the latitudes were approximately 10 times larger for the two countries, I used latitude effects that were approximately 10 times less.

Finally, GWR requires selecting a kernel and a bandwidth. As discussed in Section 4.5, the kernel is largely irrelevant; I selected the Gaussian kernel. In terms of the bandwidth, I selected an adaptive bandwidth that optimizes the leave-one-out cross-validation fit.

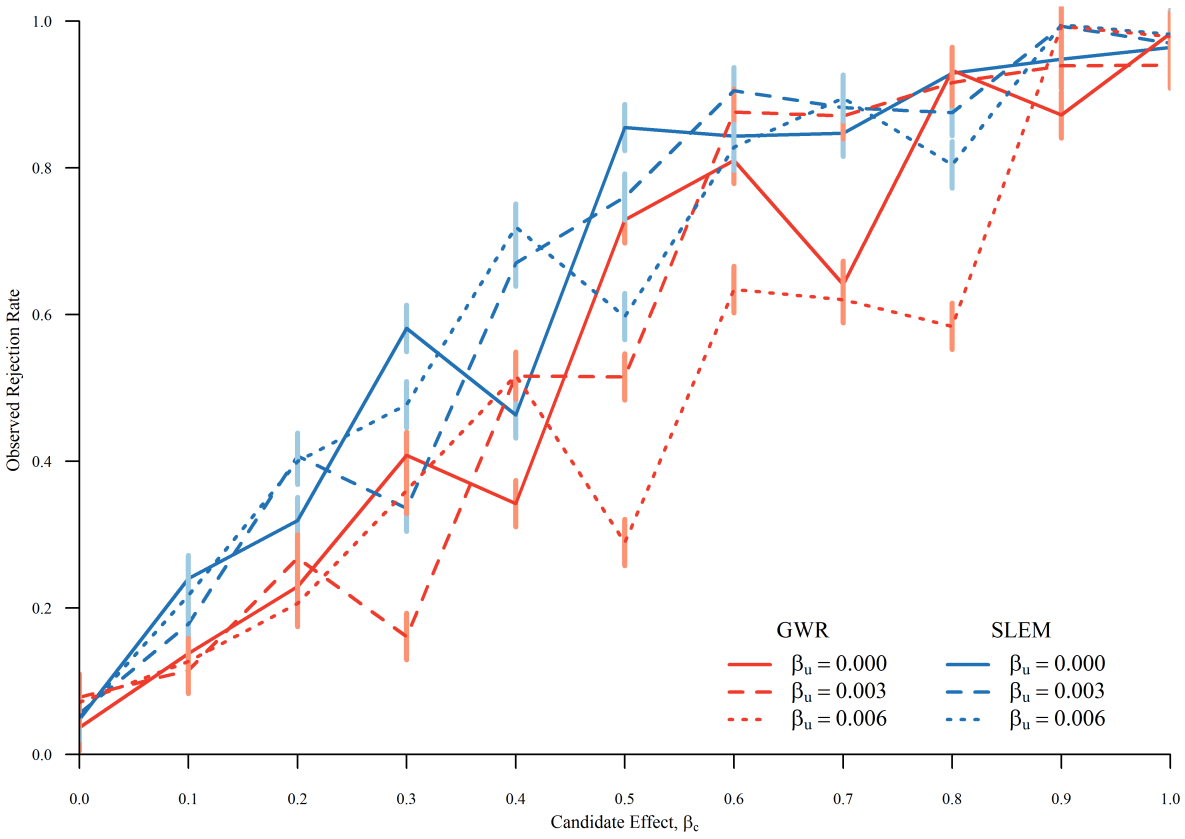


Figure 4.13: *Power curve for the effect of contamination on the rejection rate, separated by estimation method and effect of longitude (β_u). This corresponds to the regular grid map with* $\beta_v = \beta_{uv} = 0.000$.

THE RESULTS: For THE GRID, the results are very interesting in terms of power. Figure 4.13 provides the corresponding power curves as functions of the candidate effect (horizontal axis), longitudinal effect (line type), and estimation method (line color), where $\rho = 0$, $\beta_v = 0.00$, and $\beta_{uv} = 0.00$. The vertical segments correspond to 95% confidence intervals for the observed rejection rate (based on $10,000$ iterations).

Note that the SLEM method (blue lines) tends to have a higher power than the GWR method (red lines). Also note that both tend to increase as the effect of candidate support increases from 0 to 1. More importantly, note that the GWR method does not increase consistently. Similar results hold for setting β_u to 0.00 and varying β_v.

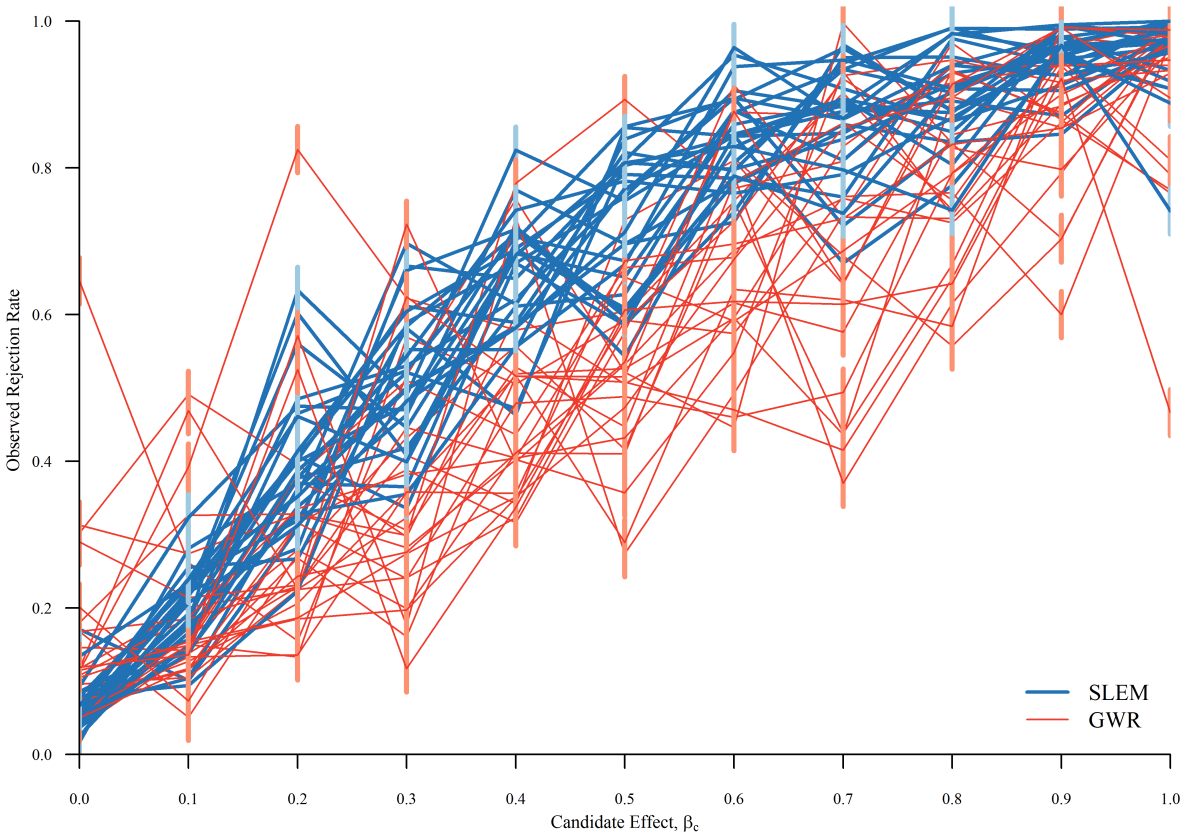


Figure 4.14: *Power curve for the effect of contamination on the rejection rate, separated by estimation method and effects of longitude and latitude and both. This corresponds to the regular grid map. Blue lines correspond to the SLEM power; red lines, GWR. Each curve corresponds to a different combination of β_u, β_v, and β_{uv}. The vertical bars are 95% confidence intervals.*

In lieu of showing many graphics telling the same story, or variations on a theme, I provide Figure 4.14. This graphic shows the various power curves for the 5×5 grid. Blue curves correspond to the power curves of SLEM; red curves, GWR. The lines correspond to varying the candidate effect under different values of β_u, β_v, and β_{uv}.

Note that the SLEM power curves, while not smooth, do tend to follow a typical power curve shape—within a band. They start low and increase with the level of contamination. When the candidate effect is 1, the power of the SLEM test is near 1.

The GWR power curves, on the other hand, do *not* follow the typical power curve shape as closely. They are incredibly variable, irregular, and erratic. They are

very dependent on the latitude, longitude, and combined effects. The power is also very dependent on the contamination level, to the point that the value of interpolating between two power estimates is quite low. Furthermore, the GWR powers tend to be less than the corresponding SLEM powers. This, and the irregular nature of the GWR power curves, suggest that SLEM is superior to GWR in this aspect and for this geographical geometry.

The next two maps differ in one particular way from this grid map: Belgium in number of divisions, Sweden in compactness. Thus, we can examine the effect of each of these two map aspects to determine how each affects the quality of the tests.

For BELGIUM, the story changes a little. Recall that the number of divisions in Belgium (11 *provincies*) was much less than that in the grid, although both were quite compact. Thus, this comparison should help us more clearly understand the effect of sample size on the quality of the two tests.

Figure 4.15 provides all of the power curves for the various combinations of positional effects for increasing candidate effect. Again, blue curves correspond to powers for the SLEM model; red curves, the GWR model. Note that there remains a lot of variation in the power for a given candidate effect. This is true for both geographic models. For instance, when the candidate effect is 1.00, the rejection rate ranges from approximately 0.25 to 0.85 for each model!

On the basis of this graphic, there is little comparison that can be made between the two methods. Both are poor. Both are irregular. Both are of low power. This suggests that the number of divisions is an important factor in the quality of the conclusions from *either* method.

Finally, let us examine SWEDEN; the story changes a bit more. Recall that Sweden had approximately the same number of divisions (*län*) as the grid (21 to 25), but was not as

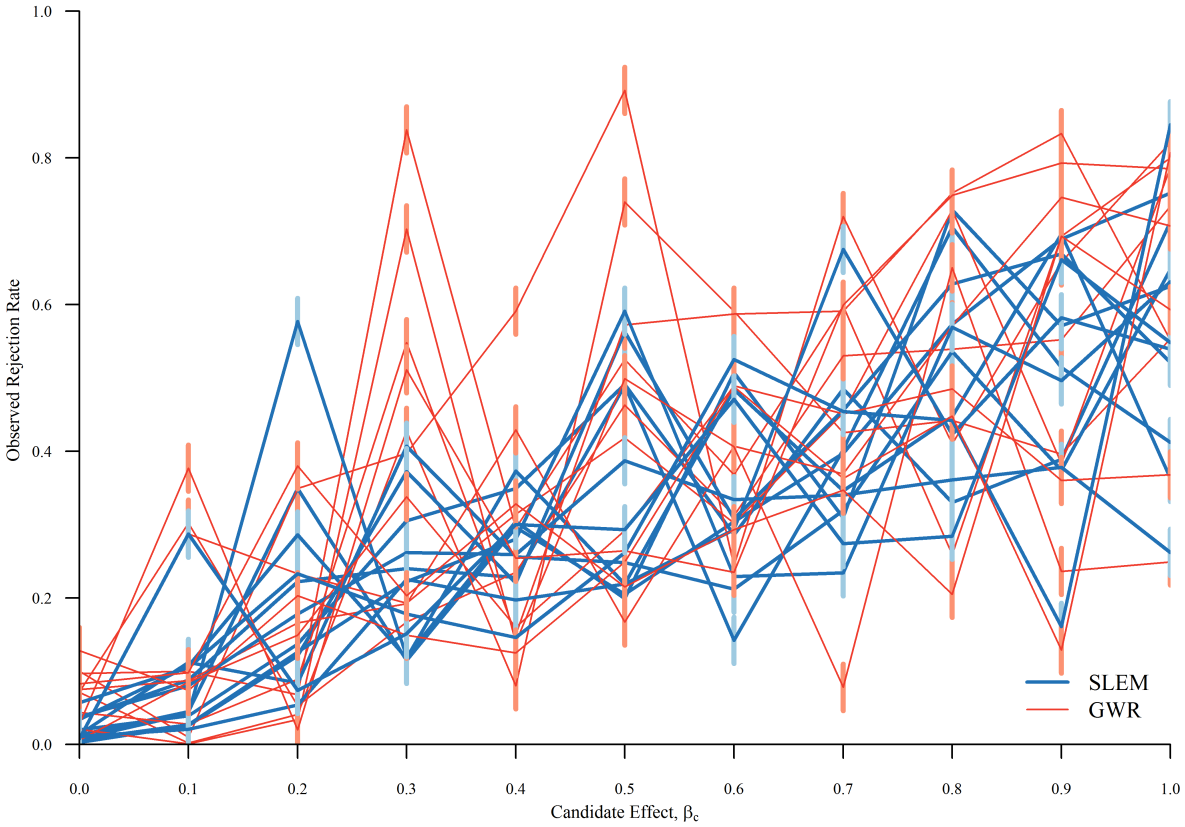


Figure 4.15: *Power curve for the effect of contamination on the rejection rate, separated by estimation method and effects of longitude and latitude and both. This corresponds to the Belgium map. Blue lines correspond to the SLEM rejection rate; red lines, GWR. Each curve corresponds to a different combination of β_u, β_v, and β_{uv}. The vertical bars are 95% confidence intervals.*

compact. Thus, this comparison examines the effect of compactness on the quality of the tests.

Figure 4.16 provides the power curves for the various combinations of positional effects for increasing candidate effect. As usual, blue lines correspond to the SLEM model; red lines, the GWR model.

Both SLEM and GWR have very low power when $\beta_{uv} \neq 0.000$. This condition corresponds to the curves that are flat until $\beta_c = 0.50$ or greater (lighter colored curves). Beyond these serious issues, the SLEM curves tend to be higher than the GWR curves.

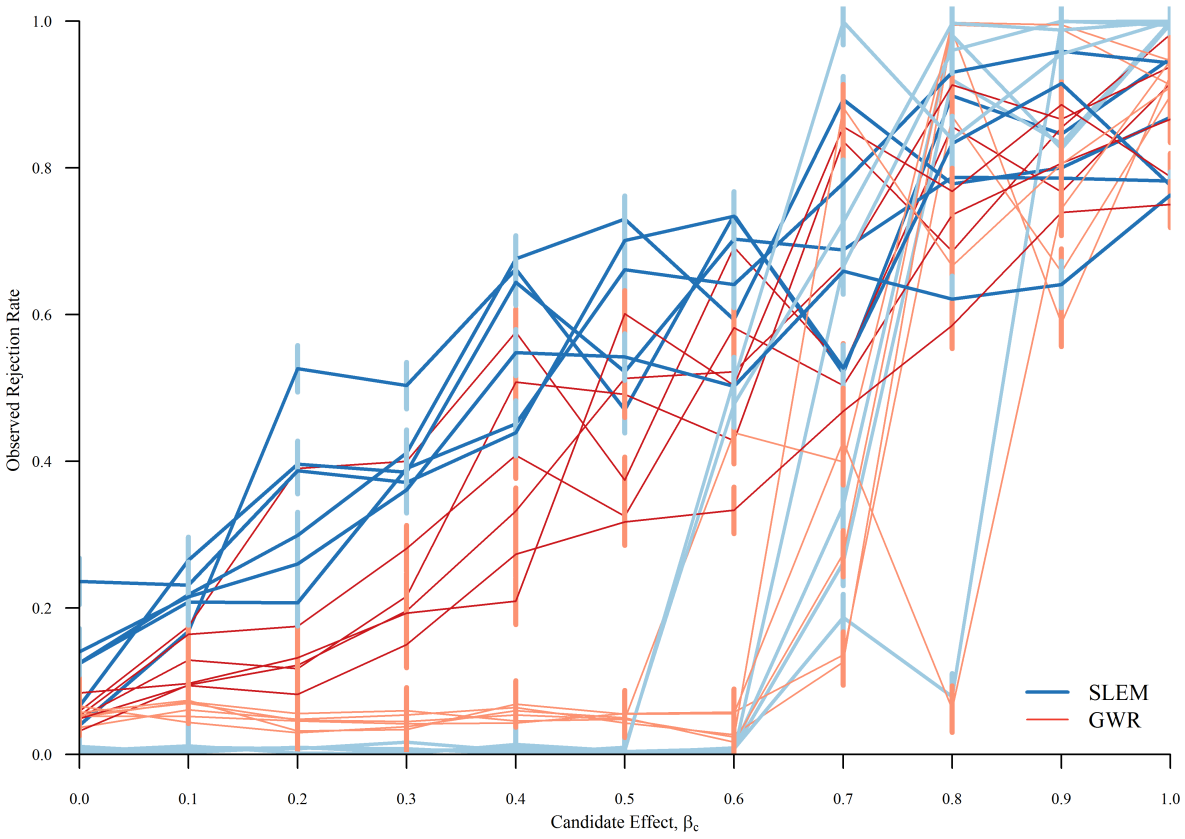


Figure 4.16: *Power curve for the effect of contamination on the rejection rate, separated by estimation method and effects of longitude and latitude and both. This corresponds to the Sweden map. Blue lines correspond to the SLEM power; red lines, GWR. Each curve corresponds to a different combination of β_u, β_v, and β_{uv}. The vertical bars are 95% confidence intervals.*

Thus, from comparing the tests' results on the grid to those on Sweden, we may surmise that the effect of compactness is minor unless there is a cross-term effect. In such cases, the power of either test is very small until the candidate effect is large.

4.8. Conclusion

Closing out the theory part of the dissertation, this chapter examined three current methods of incorporating geography in regression methods. The spatial lag model is able to reduce spatial correlation in the residuals, but it does not allow for modeling spatially-varying effects. The spatial expansion method does model spatially-varying effects, but

it does not handle contagion (neighbor) effects. The geographically weighted regression method handles all of this, but hypothesis testing is not natively supported.

The method I introduced was the spatial lag expansion method. It is a combination of the spatial lag method and the expansion method. It allows for modeling the contagion processes like the spatial lag model does, but it also allows for spatially-varying effects. Finally, it allows for parameter testing as it fits easily within the usual regression paradigm of Chapter 3. This allows one to use the appropriate regression method.

While SLEM does have a native testing paradigm, it should not be used unless theory makes a certain functional form (linear, quadratic, etc.) natural. In all other cases, to maximize power, the critical value should be estimated using simulation. In doing this, SLEM tended to outperform GWR in terms of power.

Finally, I was able to begin investigating the effect of geometry on the tests. Both tests performed better for the compact 5×5 grid than they did for Belgium, which was compact, but only had 11 divisions. For Belgium, both tests had low power and were erratic.

Both tests also performed better for the compact 5×5 grid than for the non-compact Sweden. The biggest impact happened when there were cross-positional effects. When the latitude and longitude effects were additive, the two tests performed much like they did in the case of the grid. In other words, geometry matters.

In the next two chapters, I apply all of the best methods discussed in these last three chapters to two cases. The first case is the South Sudanese Unity referendum of 2011 (Chapter 5). The second case is the 2008 US Presidential election in Colorado (Chapter 6). The former case shows clear evidence of violations of the free and fair hypothesis. The latter case does not fail any of the tests, lending credence to the assertion that the election was free and fair.

CHAPTER 5

APPLICATION: SOUTHERN SUDAN, 2011

Southern Sudan, 2011. During the colonial period, the United Kingdom ruled northern Sudan separately from southern Sudan. However, when the Republic of Sudan became independent, it was as a single, unitary entity. When it became clear that independence meant a unified Sudan, Southerners began a guerilla campaign—later a military campaign—to wrest a separate independence. The civil war lasted from 1955 to 1972 and caused almost a half million deaths (Collins 2008).

The First Sudanese Civil War ended with the Addis Ababa Accords of 1972. This peace treaty resulted in the formation of an autonomous zone in southern Sudan (Johnson 2011). During the lifetime of the treaty, tensions eased and the South was not in open rebellion. However, President Gaafar Nimeiry revoked the treaty—and the Sudanese constitution—in 1983, putting all of Sudan under Shari'a Law. Almost immediately, the South rebelled, starting the Second Sudanese Civil War.

After 22 years and almost two million dead, the government of Sudan and the leaders of the South signed the Comprehensive Peace Agreement, containing eight separate protocols designed to settle the "Southern Question." The most important aspect of this agreement was that popular referenda would be held across Sudan allowing southerners to vote on the future relationship between the North and the South (Collins 2008).

According to the Comprehensive Peace Agreement (Naivasha Agreement), southern Sudan was divided into 12 states, plus the Abyei Area. All were to have held their referenda simultaneously in January 2011. However, ambiguities in the agreement's wording allowed Sudanese President Omar al-Bashir to suspend the referendum process in the

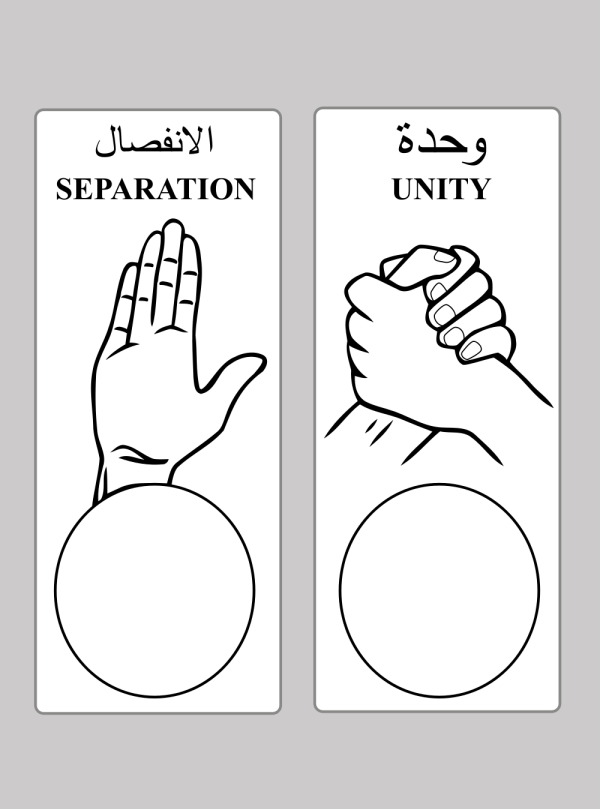


Figure 5.1: *Ballot paper for the South Sudanese unification referendum of 2011. The languages are English and Sudanese Arabic. The hand symbols are for those who know neither language. The voter is to place a thumbprint in indelible ink in the appropriate circle. The use of indelible ink also ensures people vote at most once.*

Abyei area; the governor of South Kurdufan suspended the referendum process because of personal ties to the Sudanese President; and lingering northern military action against southern rebels allowed its governor to indefinitely postpone the referendum process in Blue Nile state. Thus, January 2011 saw residents of only 10 of the 13 regions (12 southern states and Abyei) vote in the independence referendum.

As expected, the official count showed strong support for independence. Officially, 98.8% of the registered voters cast valid ballots for independence from Sudan (Sudan 2011). The support for succession at the state level ranged from 50.06% in North Kordufan (in Sudan) to 99.98% in Unity (in South Sudan).

Election observers included former US President James Carter, former UN Secretary General Kofi Annan, and current actor George Clooney. The Carter Center group

visited both President al-Bashir and First Vice President Salva Kiir (the future president of the Republic of South Sudan). The group visited several polling stations. Carter (2011) reported

> Before leaving Sudan early on January 16, voter turnout had exceeded 90 percent in the South and 40 percent in the North among the 618 sites visited by our observers. A few sites in S. Sudan had 100 percent turnout and were practically unanimous for independence. The entire exercise was orderly, pleasant, and productive, and it is expected that the official returns will lead to a new nation, and that The Carter Center will remain involved in both countries in promoting peace, democracy, and better health and education.

However, even without charges of fraud and reports of irregularities, certain facts raise questions. Ten counties reported no votes for unity, and 21 reported fewer than five. The average number of votes in these 21 counties was over 35,000. Furthermore, 12 counties reported no invalidated ballots; that is, the government stated that the voters in those 12 counties all filled out their ballots perfectly. The remaining 91 electoral divisions averaged invalidation counts of just 68. Most interestingly, the invalidation rate differed significantly between the northern voters (3.2%), the out-of-country voters (0.3%), and the southern voters (0.1%).

On its face, these results are not compatible with the 'Free and Fair' hypothesis. And yet, the US Ambassador to the United Nations (Rice 2011) did not question the vote's outcome

> On behalf of the people of the United States, let me again congratulate the people of South Sudan for a successful and historic referendum in which the overwhelming majority of voters chose independence.

Nor did she question the vote's legitimacy.

5.1. Introduction

The previous three chapters provided several tests of the free and fair hypothesis, each centered on a given level of data provided. In this election, the government of South Sudan provided much information: counts of invalidations, blank ballots, referendum support, and total ballots. The Global Administrative Areas database also provides shapefiles, allowing geographic analysis (GADM 2014b). Unfortunately, the division names in the voting records do not match the third-level divisions in the shapefiles. Thus, just for the geographic tests, the data are aggregated to the 10 second-level administrative divisions (states).

5.2. Digit Test

In Chapter 2, I covered several options for improving upon the current Benford test. None were good. The empirical Benford tests both suffered from low power, thus making them almost useless. The generalized Benford tests had very high power, which meant they easily rejected the fair hypothesis. Of all five tests, I weakly suggested using the Likelihood Simulation test as a gatekeeper. If the election passed that test, then there was no evidence of count tampering; if it failed, there was little evidence of it.

Thus, I perform the Likelihood Simulation test on the full election returns. Figure 5.2 provides a histogram of the simulated log-likelihoods, with the observed log-likelihood, -223.9304, indicated. This value corresponds to an approximate p-value of 0.0012. This is much less than the usual $\alpha = 0.05$. As such, we cannot conclude that there is no evidence of people manually changing the election counts. While this is not strong evidence, it does raise a flag in this election. Let us keep it in mind.

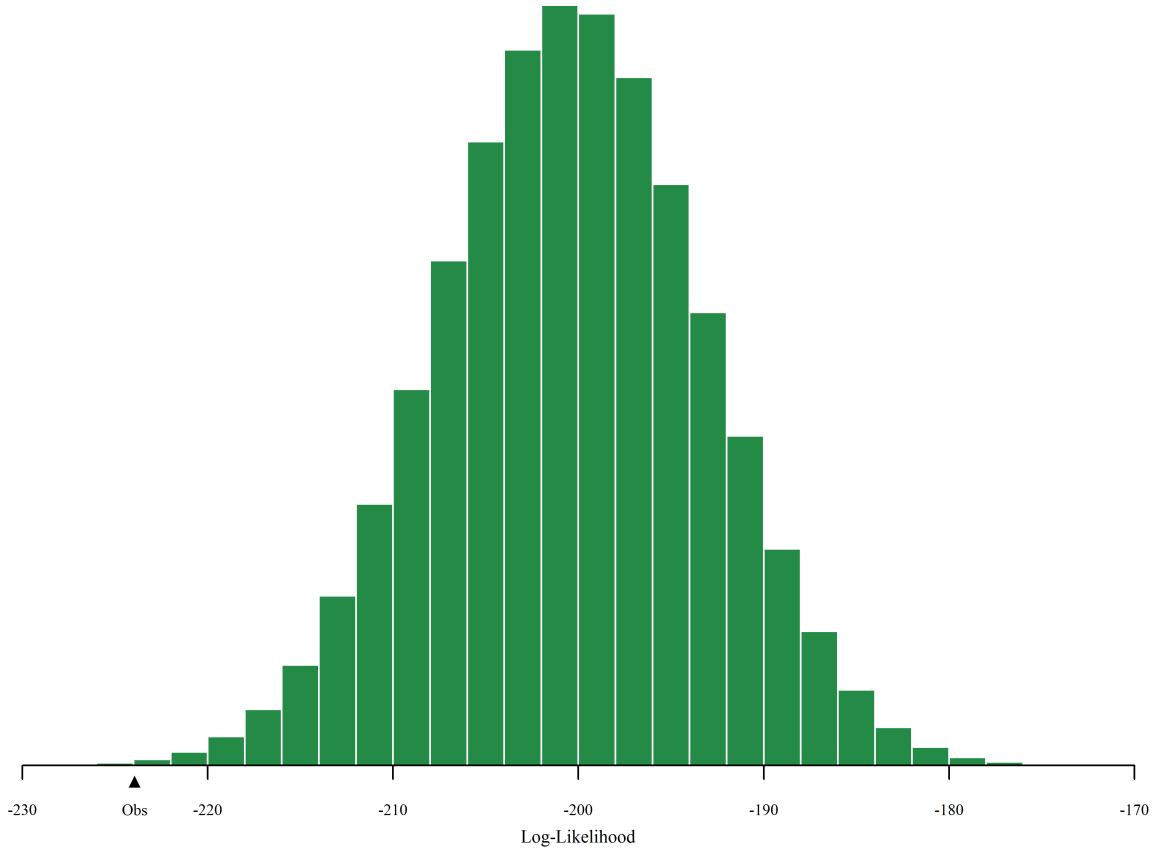


Figure 5.2: *Distribution of the log-likelihood for the 2011 South Sudanese unity referendum. The observed test statistic,* -223.9 *is indicated on the graphic.*

5.3. Regression Tests

In Chapter 3, I examined two complementary aspects of testing. The first concerned the appropriate least squares regression method. I concluded that feasible generalized least squares (FGLS) was the best, while noting that weighted least squares (WLS) would be similar in terms of estimation and would tend to be slightly faster.

The second aspect concerned estimating division membership into two populations: those that were fair and those that were not. Here, I concluded that the grid search was better than the Bayesian method in terms of mean square error and time. The drawback is that the usual test statistic no longer follows its expected distribution; the distribution must be simulated.

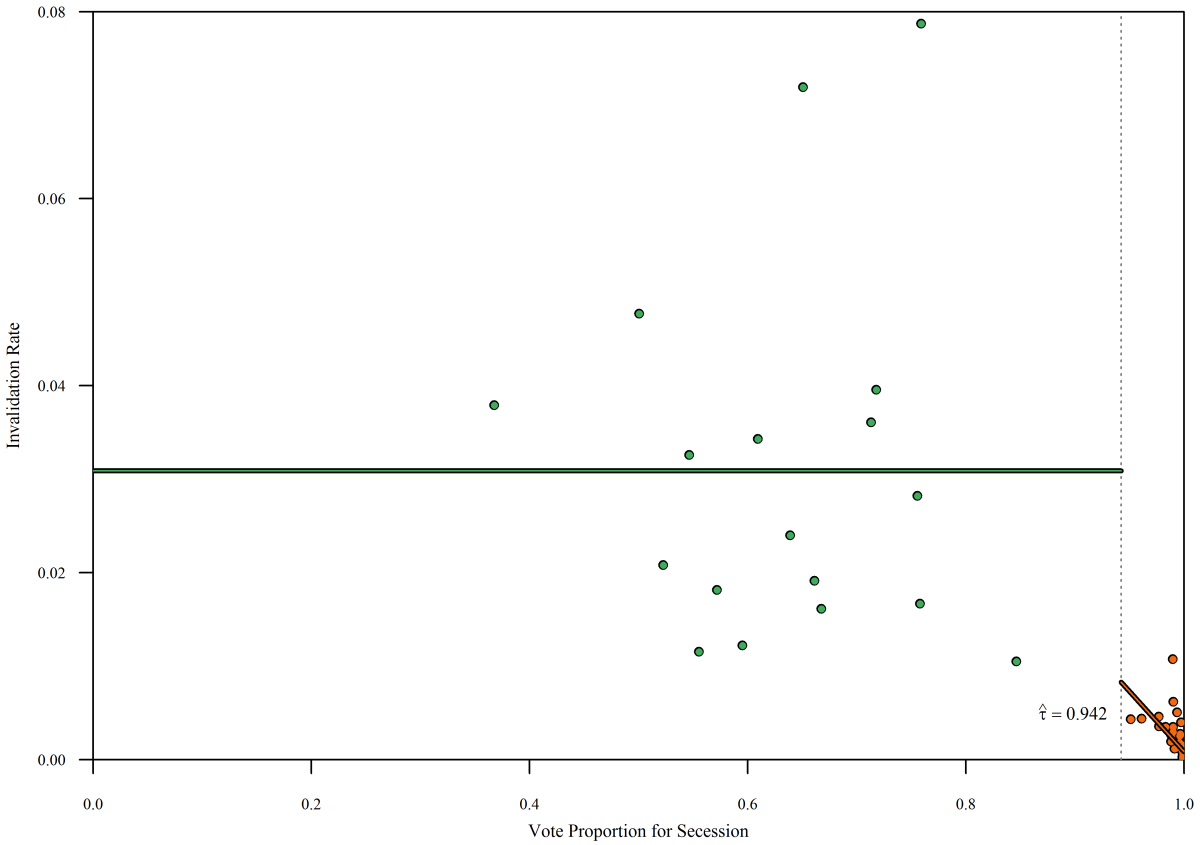


Figure 5.3: *Invalidation plot for the 2011 South Sudanese referendum. The optimal threshold is* $\hat{\tau} = 0.942$.

Figure 5.3 is the invalidation plot for this election. The grid search produced an optimal threshold of $\hat{\tau} = 0.942$. The estimated candidate effect for those divisions with independence support greater than 0.942 is -0.133, which is also practically significant. The observed t-value is -5.938, which would correspond to a p-value of 1.42×10^{-7} if the test statistic followed the usual distribution. Simulation of the test statistic under the null hypothesis gives an estimated central 95% confidence interval of -2.78 to 2.74 and an estimated p-value of 0.0013.

Thus, based on this test, we can conclude that there is significant evidence of unfairness in this election. The data support the contention that those voting in favor of independence had a lower probability of having their ballots invalidated than those voting in favor.

5.4. Geography Tests

Finally, Chapter 4 covered the effects of geography and how best to handle it. The current 'gold standard,' geographically weighted regression (GWR), performed well in comparison to my suggestion of spatial lag expansion method (SLEM). Neither performed well when the number of divisions was low (e.g., Belgium). Both performed well when the number of divisions was high and the map was compact (e.g., the grid).

South Sudan is more like Belgium than the grid. Because of data-matching issues, I must aggregate to the state level, and South Sudan has only 10 states.

Using 10,000 iterations, I estimated the critical value of the GWR test to be 0.026 and of the SLEM test to be 31.46. These led to estimated p-values of 0.0002 from GWR and 0.0268 from SLEM.

Figure 5.4 provides maps of the candidate effects estimated by the GWR method; Figure 5.5, the SLEM method. In both cases, all estimated effects are negative, which echoes the conclusions of Section 5.3.

The GWR method estimates that the candidate effect ranges from -0.102 to -0.095. Thus, GWR does not suggest a strongly spatially varying candidate effect. On the other hand, the SLEM method estimates that the candidate effect ranges from -0.255 to -0.071, which does suggest a strong effect of space on the candidate effect. Note that Section 5.3 estimated the candidate effect to be -0.133, which is not within the GWR range but is within the SLEM range.

In either case, there is statistically significant evidence for a violation of the free and fair hypothesis in this election ($p_{\text{GWR}} = 0.0002; p_{\text{SLEM}} = 0.0268$).

5.5. Using Additional Information

In Section 5.3, I used the general techniques discussed in Chapter 3. This election, however, has additional information that should not be discarded.

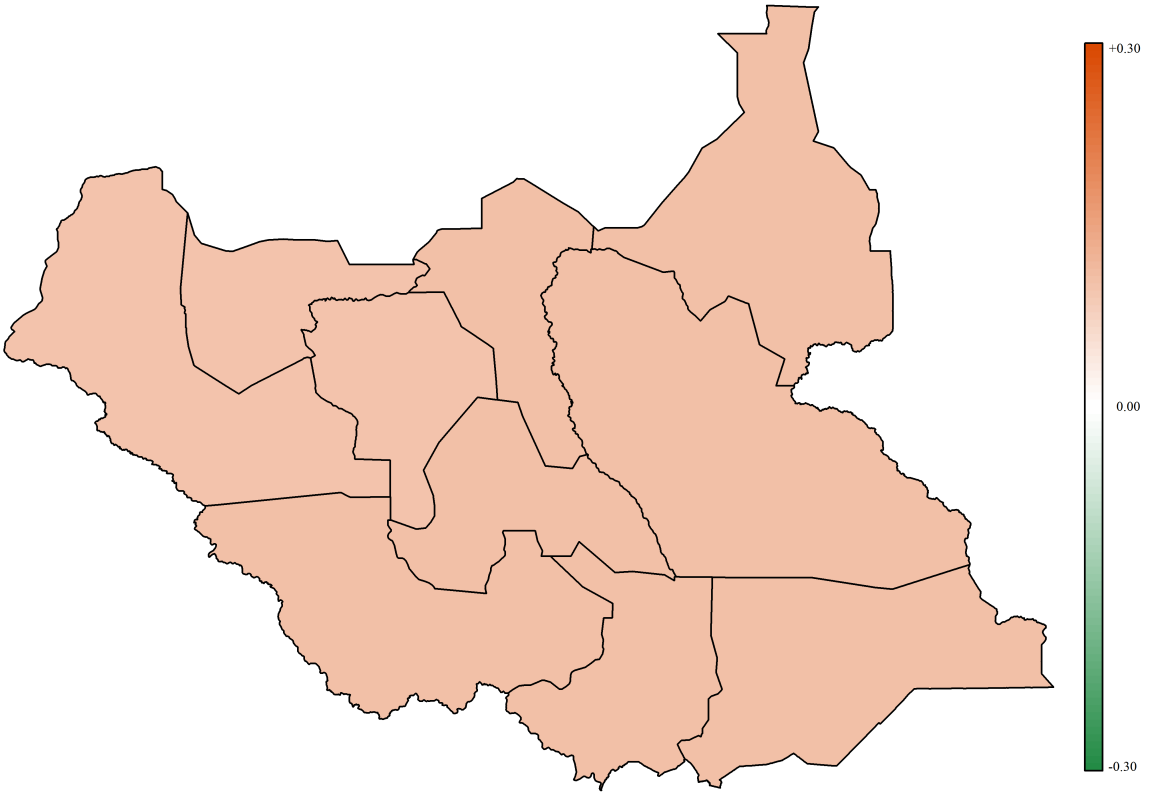


Figure 5.4: *Maps of the candidate effects for the GWR method. Darker oranges correspond to higher (negative) candidate effects. Shades are equivalent across this and Figure 5.5.*

The South Sudanese referendum had citizens cast ballots across the world. In foreign countries, the ballots were cast in Sudanese consulates and forwarded to the elections commission in Juba, South Sudan, for counting. In northern Sudan, the ballots cast in the electoral divisions were counted *in situ*. Those ballots cast in South Sudan were counted in South Sudan. Thus, there are potentially two populations: ballots counted in Sudan and ballots counted in South Sudan.

Figure 5.6 provides the invalidation plot for this partitioning. Notice that, with one exception (Raga County in Bahr el Ghazal state), those ballots counted in Sudan correspond to those in the left section of Figure 5.3. Thus, the "blind" grid search was able to identify that these ballots were counted under a different process than the others.

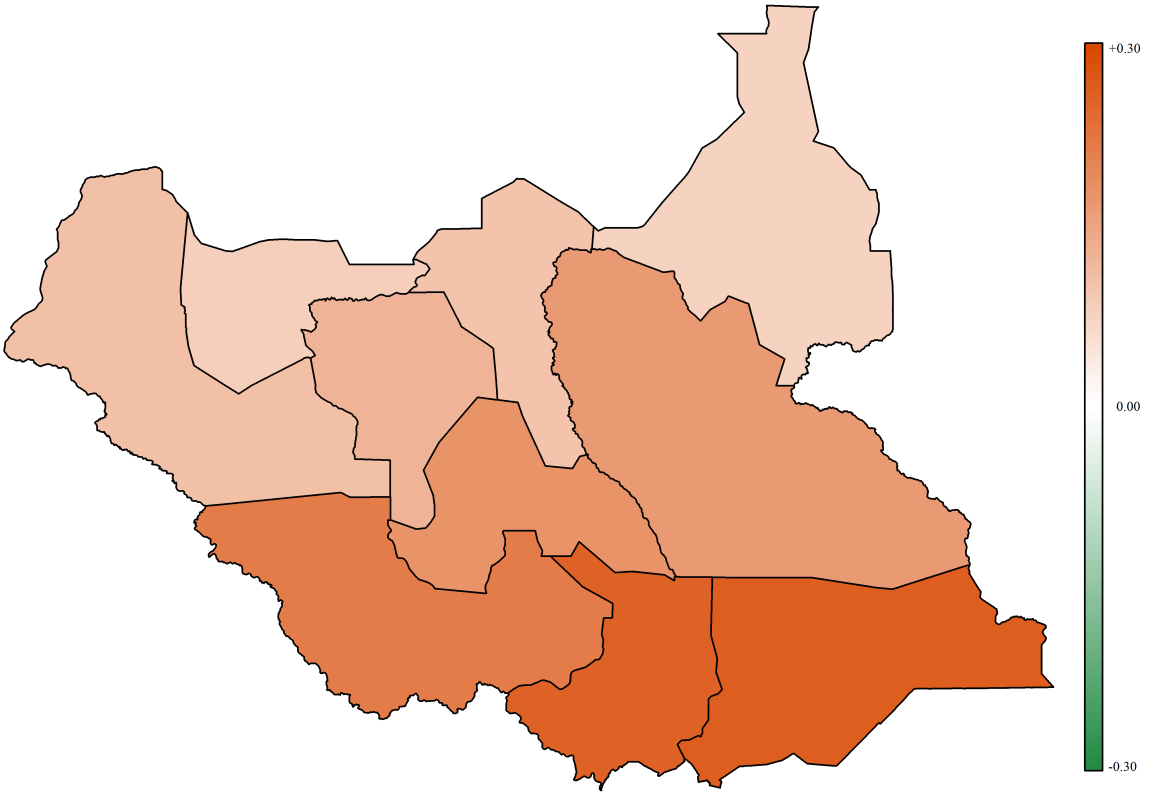


Figure 5.5: *Maps of the candidate effects for the SLEM method. Darker oranges correspond to higher (negative) candidate effects. Shades are equivalent across this and Figure 5.4.*

For the record, the candidate effect for the Sudanese-counted ballots (green) is 0.03661, with a p-value of 0.421; for the South Sudanese-counted ballots, -0.079360 ($p < 0.0001$). Thus, we again have evidence that the election was not fair. Now, however, we can point to a problem with the ballots counted in the southern divisions.

5.6. Conclusion

In this chapter, I illustrated how to implement the tests of the previous three chapters in a real election. In 2011, the self-identified citizens of South Sudan went to the polls to voice their hopes for an independent future.

The government of South Sudan reported that 3,769,350 voted in favor of independence and 44,877 voted for continuing the current political situation with Sudan.

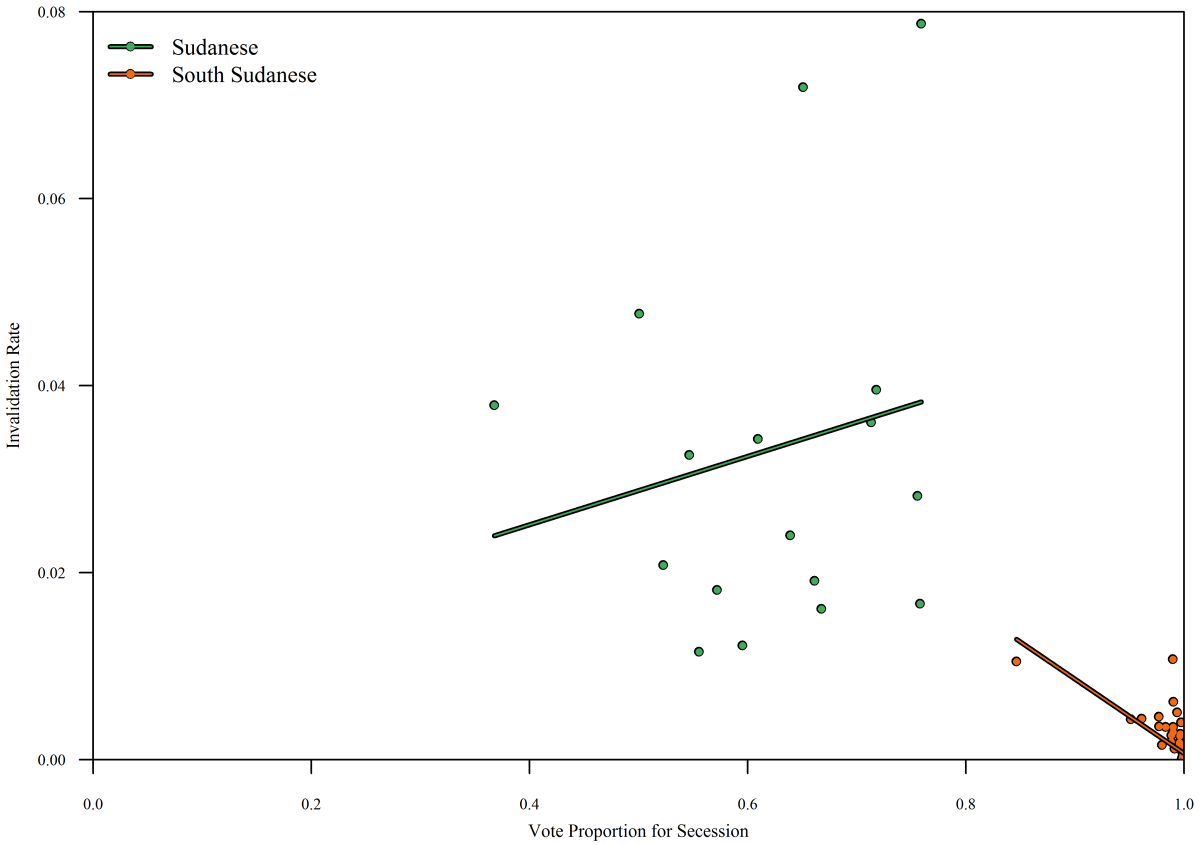


Figure 5.6: *Invalidation plot for the 2011 South Sudanese referendum. The divisions are colored based on where the ballots were counted. Those in green were counted in Sudan; in orange, in South Sudan.*

Twenty-one counties reported fewer than five invalidated ballots, raising concerns about the fairness of the vote.

The Likelihood Simulation method, testing the reported vote counts for independence, concluded that the reported results did not meet expectations. However, recall from Chapter 2 that even failing this test is not strong evidence against the claim of fairness. It is, at most, a red flag that vote counts may have been manually adjusted.

Weighted least squares on two populations provided that strong evidence, however. Not only was the candidate effect statistically significant ($p = 0.0013$), but it was substantively significant ($\beta_c = -0.13$).

Finally, the two geographical methods both concluded that there was a statistically significant candidate effect. They did, however, disagree on whether that effect varied

across the map or was essentially constant: SLEM concluded it varied spatially from -0.255 to -0.071, GWR did not.

Thus, taking these results as a whole, I conclude that this election was not fair to those voting for unity with Sudan. I cannot conclude that this inequality was due to election fraud or to an unfair electoral system. I can say, however, that there is no evidence that more people voted for unity than for independence. In other words, the election was unfair, but the results appeared to reflect the general will of the people.

CHAPTER 6

APPLICATION: COLORADO, 2008

Colorado, 2008. The United States holds presidential elections every four years, on years evenly divisible by four. The president is elected by the Electoral College, which is comprised of 538 members selected from the 51 federal members. Strictly speaking, those 51 members (50 states and the District of Columbia) are not all states, but I shall refer to them as such. The size of the state delegation equals the size of its federal delegation (Article II). The size of the delegation from the District of Columbia is set by the US Constitution at 3 (Amendment 23).

The Electors vote at their individual statehouses on the Monday after the second Wednesday of December, (December 15, 2008, in this election) and forward their votes (one for president and one for vice-president) to the President of the US Senate, who brings them to the joint session of the US Congress. Four tellers, one from each party in each chamber, count them before the new members of Congress (3 USC 1).

The voters chose these Electors on the Tuesday after the first Monday of November. In all states, the Electors are presented as a slate. In 49 states, a single slate of Electors win the state. Maine and Nebraska use the Massachusetts method: two electors are chosen at the state level and one from each Congressional district. Thus, it is the 51 state-level and 5 Congressional District-level jurisdictions that hold the elections, not the unitary federal system.

The 2008 US Presidential election pitted Republican Senator John McCain against Democratic Senator Barack Obama. The incumbent president was George W. Bush, who belonged to the same party as Senator McCain. Due to several missteps in the McCain

campaign and to the fact that solidly Democratic states controlled over 300 electoral votes, the winner of the Presidential election was known well before the ballots were cast.

The US state of Colorado is the eighth largest in terms of area; the 22nd in terms of population. While it is termed a "purple mountain" state in reference to its balance of Democratic (blue) and Republican (red) voters, the opinion polls in Colorado had Obama leading McCain by 7–10%, thus polling analysts forecasted this purple mountain state blue.

November 4th came and went. Senator Obama carried Colorado by almost 9%. Senator McCain did not call for an investigation, which would have been tantamount to declaring electoral fraud had taken place. There were no protests, no riots, no deaths due to the election. Compared to the other elections discussed in this dissertation, it was boring. As such, there is no expectation of detecting fraud here.

However, there is interesting structure to the results—a structure that hints at election-day problems.

6.1. Introduction

In the previous chapter, I examined the 2011 Unity Referendum of South Sudan using the techniques discussed in the first four chapters of this dissertation. Because of the dearth of information, there was little to be done beyond examining the relationship between the invalidation rate and independence support.

Colorado offers much more information. While I still apply the basic tests discussed in Chapters 2 to 4, I follow up using the additional available demographic information. Since Colorado is one of 50 states in the United States of America, a census was taken there in 2010. Thus, that demographic information is relatively recent and useful.

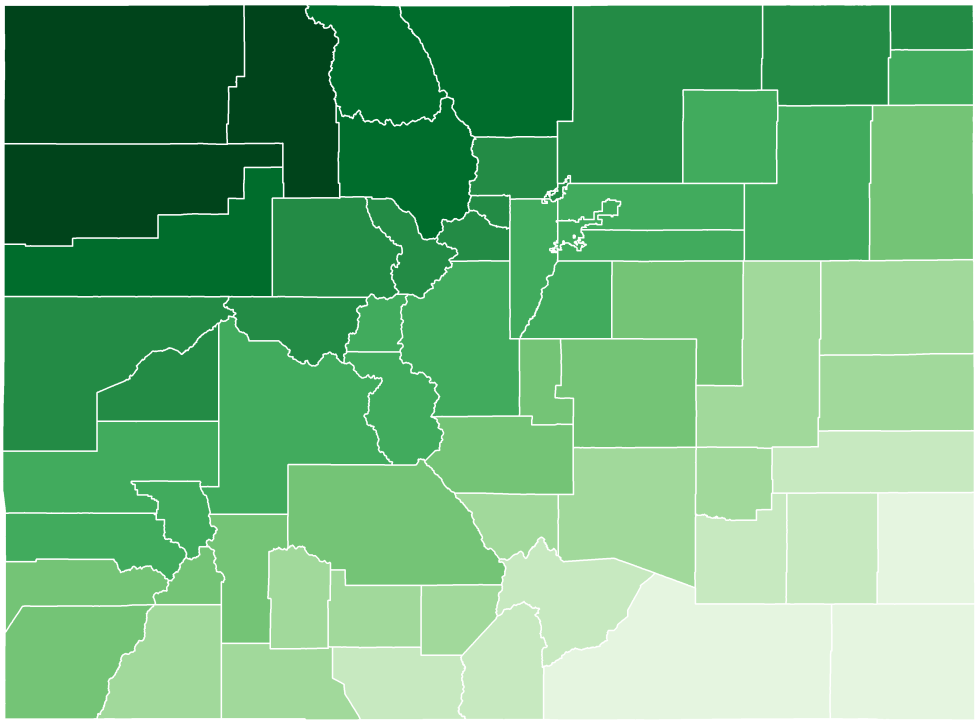


Figure 6.1: *A map of the US state of Colorado. Note that the number of divisions (n=65) and its relative compactness makes is similar to the grid of Chapter 4.*

6.2. Digit Test

Recall Chapter 2 in which I covered several options for improving upon the current Benford test—none were excellent. The empirical Benford tests both suffered from low power, and the generalized Benford tests suffered from high sensitivity to the null distribution. Of all tests, I weakly suggested using the Likelihood Simulation test as a gatekeeper. If the election passed that test, then there was no evidence of count tampering; if it failed, there was little evidence of it.

Thus, I perform the Likelihood Simulation test on the full election returns. Figure 6.2 provides a histogram of the simulated log-likelihoods, with the observed log-likelihood, -128.9435, indicated. This value corresponds to an estimated p-value of 0.633 on one-million iterations. This is greater than the usual $\alpha = 0.05$. As such, there is no evidence of people manually changing the election counts.

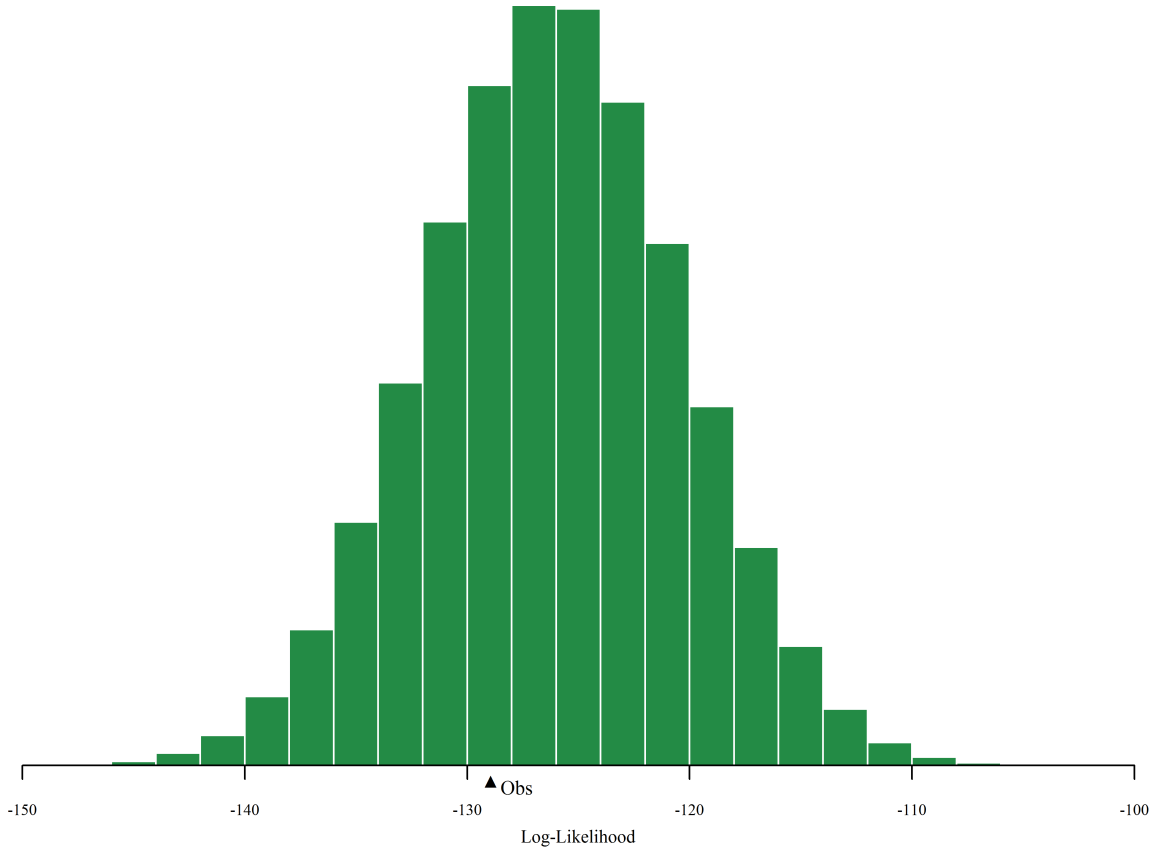


Figure 6.2: *Distribution of the log-likelihood for the 2008 Coloradoan presidential election. The observed test statistic,* -128.9 *is indicated on the graphic.*

This is not a surprise. Before election day, Senator Obama was predicted to win Colorado by a margin of between 7 and 10%. Thus, even had there been the ability to carry out a systematic fraud campaign, it would have had to be of such magnitude to make it obvious.

6.3. Regression Test

In Chapter 3, I concluded that using weighted least squares regression in combination with a grid search for the optimal breakpoint would allow one to better test if the invalidation rate was independent of the candidate support rate.

Figure 6.3 is the invalidation plot for this election. The grid search produced an optimal threshold of $\hat{\tau} = 0.716$. The estimated candidate effect for those divisions

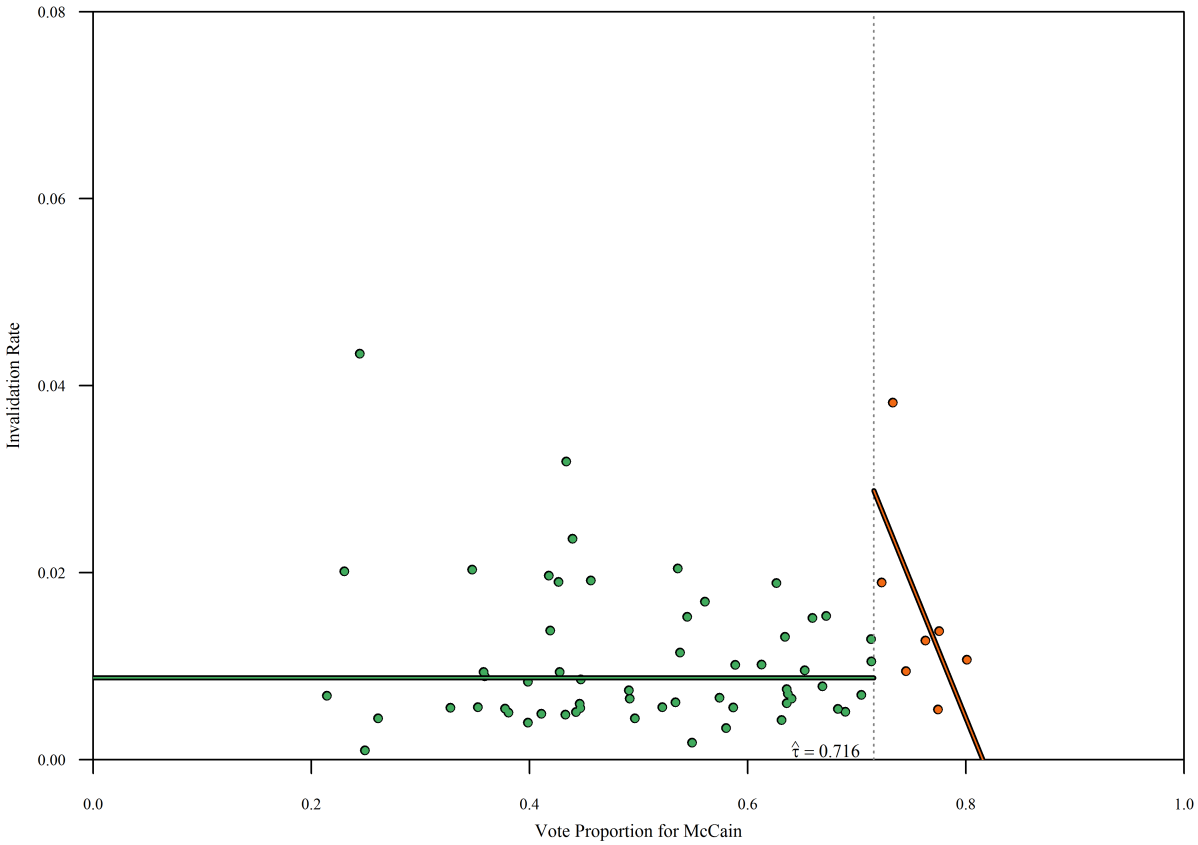


Figure 6.3: *Invalidation plot for the 2008 Coloradoan presidential election. The optimal threshold is* $\hat{\tau} = 0.716$.

with independence support greater than 0.716 is -0.2302, which is practically significant. The observed t-value is -1.601, which would correspond to a p-value of 0.170 if the test statistic followed the usual distribution.

Figure 6.4 is a histogram of the simulated test statistics under the null hypothesis. This produces a central 95% confidence interval of -2.74 to 2.79 and an estimated p-value of 0.275, which is much greater than the usual $\alpha = 0.05$.

Thus, based on this test, we cannot conclude that there is significant evidence of unfairness in this election. The inherent randomness of election returns is likely the reason behind the steep slope of the right line. Again, as with the digit test above, I had no expectation of this election showing signs of electoral tampering or of political unfairness.

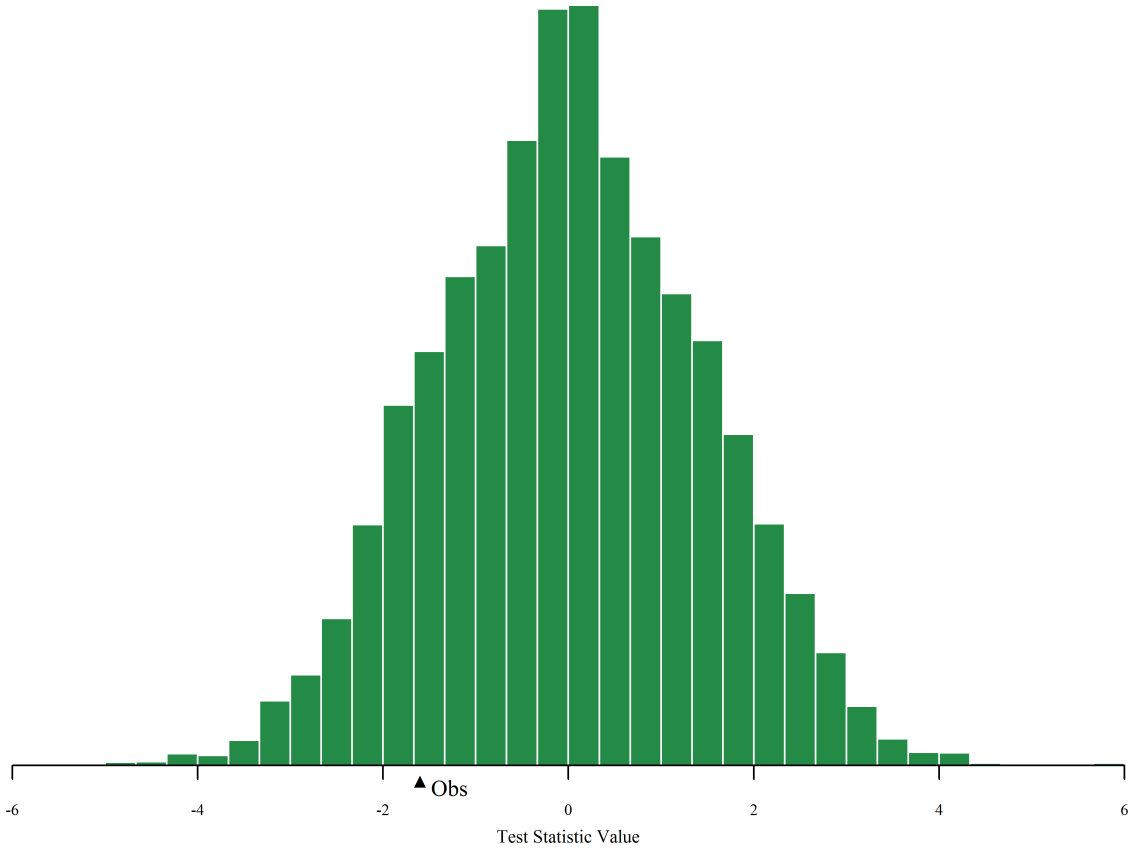


Figure 6.4: *The estimated distribution of the test statistic for the Colorado election, with the observed value shown. A 95% confidence interval is from* -2.74 *to* 2.79 *on 10,000 iterations.*

6.4. Geography

Next, Chapter 4 covered the effects of geography and how best to handle it. The current 'gold standard,' geographically weighted regression (GWR), performed well in comparison to my suggestion of spatial lag expansion method (SLEM). Neither performed well when the number of divisions was low (e.g., Belgium). Both performed well when the number of divisions was high and the map was compact (e.g., the grid). Colorado fits into this latter group. As such, I would expect both tests to be of high power, with the SLEM test being slightly more powerful than the GWR test.

Figure 6.5 shows the invalidation rate for each of Colorado's 66 counties (GeoCommons 2014). The rate ranges from 0.10% in Pitkin County (west central) to 4.34%

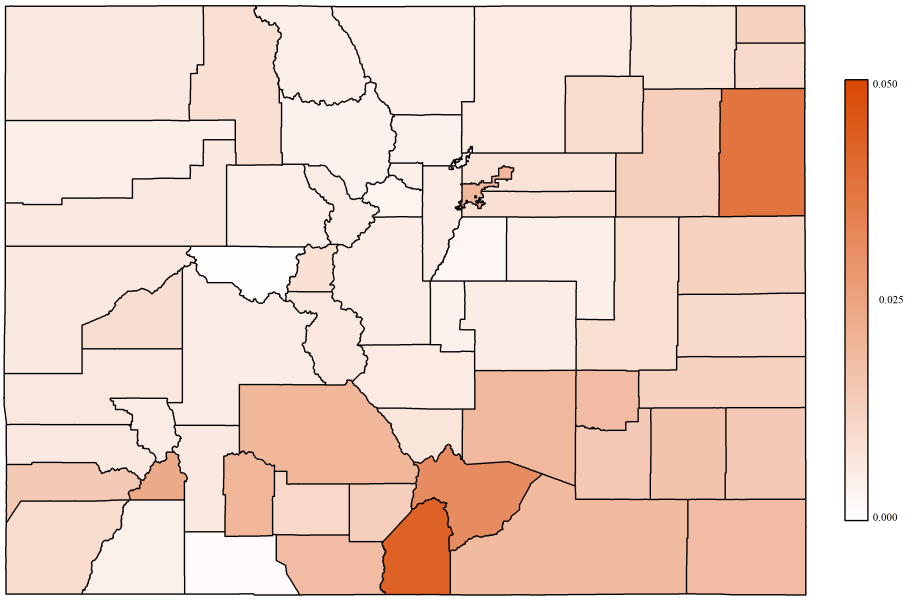


Figure 6.5: *A map of the observed invalidation rate for Colorado in the 2008 Presidential election. Note the hot spots in the east and south.*

in Costilla County (central south). In addition to the southern area, the northern eastern area also has a higher than average invalidation rate, with a 3.82% rejection rate in Yuma County. These findings suggest that geography may play a role here, that effects are spatially varying in this election.

To determine if, and to what extent, geography is a factor, I performed the recommended tests of Chapter 4. As the number of divisions was not small, I used the interaction model for SLEM; that is, I used latitude, longitude, and their product. Using 10,000 iterations, I estimated the critical value of the GWR test to be 2.45×10^{-9}, and of the SLEM test to be 12.575. The observed test statistics were 1.10×10^{-7} and 8.867, respectively. The resulting estimated p-values were 0.1365 and 0.1868. Neither of these is less than the usual $\alpha = 0.05$.

Thus, there is no evidence that the invalidation rate pattern is anything other than random noise. Figure 6.6 provides maps of the candidate effects estimated by the two methods. The left panel maps the effects estimated using GWR; the right, SLEM.

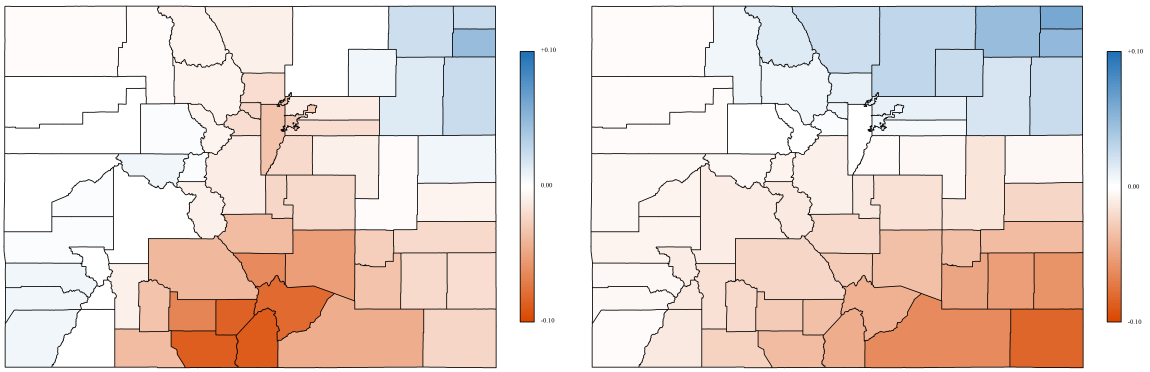


Figure 6.6: *Maps of the candidate effects for the two methods. The GWR results are the left map; SLEM, right map. Darker oranges correspond to higher negative candidate effects, blue to higher positive effects. Shades are equivalent across maps, ranging from* $+0.10$ *to* -0.10*.*

6.5. More Geography

As a part of the United States of America, Colorado had a census taken in 2010, just two years after this election. Beyond a simple counting, the census estimates demographic information. Thus, to extend these methods, I will also include demographic information to further test if the election was fair to all tested groups.

As usual, the dependent variable for the model is the proportion of votes invalidated in the county (Figure 6.5). However, because of the amount of data available, the candidate support rate is not the only independent variable. Initially, the independent variables are the proportion of the vote for McCain and several demographic variables: turnout rate, percent older than 65, percent female, percent white, percent black, percent Hispanic, percent speaking English as a non-native language, percent with High School Diploma, the percent with Bachelors Degree, and the poverty rate. All of these variables are measured at the county level.

Because of multicollinearity, many of these variables are excluded from the final model. Table 6.1 provides summary statistics of these variables, including three measures of global spatial correlation (Anselin 1995; Getis and Ord 1997). That the independent variables are spatially varying, and that the dependent variable is spatially varying sug-

	Mean	St Dev	I	C	G	Normality
Invalidation percent	1.10	0.81	0.370*	0.568*	0.091*	<0.001
Vote for Obama	45.25	15.34	0.357*	0.582*	0.088*	0.1186
Vote for McCain	52.85	15.28	0.363*	0.577*	0.078	0.1155
Turnout rate	89.64	3.66	0.119*	0.884	0.080	0.0100
White	93.18	4.00	0.248*	0.697*	0.080	<0.001
Black	1.72	2.37	0.216*	0.781	0.125*	<0.001
Hispanic	19.21	14.30	0.399*	0.586*	0.100*	<0.001
English not first language	14.15	9.92	0.209*	0.748*	0.091*	<0.001
Percent graduating high school	88.14	6.23	0.277*	0.721*	0.080	0.0181
Percent with bachelors degree	28.80	12.42	0.343*	0.637*	0.089*	<0.001
Poverty rate	12.86	5.57	0.428*	0.578*	0.084	0.0842
Elderly rate	14.88	4.88	0.300*	0.637*	0.079	0.0814
Female	48.06	3.74	-0.063	1.001	0.079	<0.001

Table 6.1: *Univariate summary statistics of the variables under consideration in this research. The three measures of global spatial correlation are Moran's I, Geary's C, and Getis and Ord's G. Stars represent statistical significance at the* $\alpha = 0.05$ *level. The value for Normality is the p-value associated with the Shapiro-Wilk test.*

gests that geography *may* matter in finding relationships between the dependent and the independent variables. Regardless, the effect of the parameters may be spatially-varying.

Before settling on a model, one needs to examine multicollinearity in the independent variables. Not surprisingly, significant multicollinearity exists in several variables. Percents black and white are highly correlated. As such, I arbitrarily included percent black and excluded percent white. That Hispanic and English as a non-native language rates were correlated is not surprising (I retained English as a non-native language in the regression). However, the education variables were also collinear with these two (neither included). McCain support and Obama support were highly correlated, so I included just the McCain support (the Republicans were the incumbent party at the national level).

Thus, the final model includes the usual dependent (invalidation rate) and independent (McCain support) variables. Also included are turnout rate, percent black, percent non-native English speaker, poverty rate, percent elderly, and percent female. If the effect of any of these seven variables is statistically significant at the county level, there is evidence of some level and type of unfairness in this election.

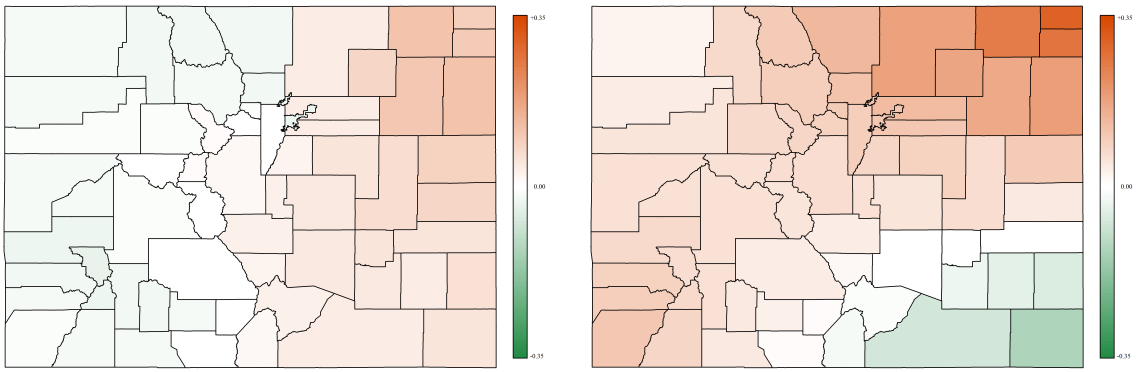


Figure 6.7: *Maps of the effects of turnout on the invalidation rate for the two methods. The GWR results are the left map; SLEM, right map. Darker oranges correspond to higher positive effects, green to higher negative effects. Shades are equivalent across maps, ranging from* $+0.10$ *to* -0.10*.*

As a whole, the GWR model is not significantly different from the null model using only neighborhood average and geographical position as predictors ($p = 0.635$), but the SLEM model *is* ($p = 0.0003$). This indicates that the SLEM model detected demographic or political effects that GWR did not.

Testing for statistical significance is a matter of calculating the test statistic (difference in log-likelihoods) for each variable against the model omitting that variable. The distribution of these test statistics can be estimated using simulation, which I did using 10,000 iterations.

The following sections cover only three of the variables, as I found these the most interesting. As the GWR model was not significantly different from the null model, I will not discuss the statistical significance of any of the variables. As the SLEM model *was* statistically significant, I do discuss the significance of the three variables below.

6.5.1 Turnout. Let us first look at the effect of turnout on the invalidation rate, controlling for all other variables discussed above. The two maps are given in Figure 6.7, with the left map being the GWR estimates and the right map being the SLEM estimates ($p = 0.0436$). The SLEM method indicates a much higher effect of turnout on

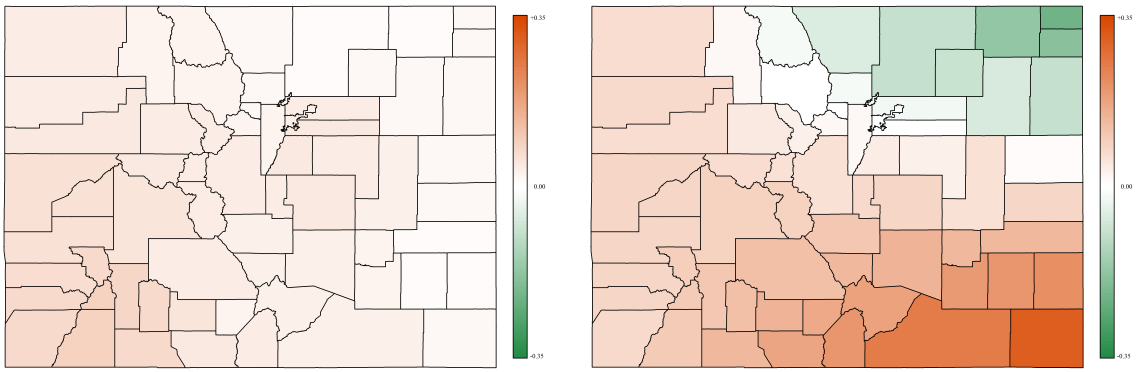


Figure 6.8: *Maps of the effects of the poverty rate on the invalidation rate for the two methods. The GWR results are the left map; SLEM, right map. Darker oranges correspond to higher positive effects, green to higher negative effects. Shades are equivalent across maps, ranging from* $+0.35$ *to* -0.35*.*

the invalidation rate, 0.298; GWR predicts the highest effect is only 0.115. Both suggest a high effect in the northeastern corner of the state, which actually had above-average turnout. This suggests the vote counters may have suffered from fatigue. This observation explains the higher-than-average invalidation rate in Yuma County (see Figure 6.5).

6.5.2 Poverty. In terms of the effect of the poverty rate on the invalidation rate, the SLEM model predicts much higher effect than the GWR model (Figure 6.8). The highest effect predicted by GWR is 0.081, whereas the highest effect predicted by SLEM is 0.302 ($p = 0.0019$). The areas of intensity also differ across the two models. GWR has the highest effects in the west, while SLEM has them in the east: northeast with the highest negative effect, southeast with the highest positive effect.

6.5.3 Candidate Effect. Finally, let us examine the candidate effect, controlling for these other variables (Figure 6.9). Again, note that SLEM predicts higher effects than does GWR. Both predict high positive effect in the northeast and high positive effect in the south. However, the SLEM effect is not statistically significant at the usual $\alpha = 0.05$ level ($p = 0.0509$).

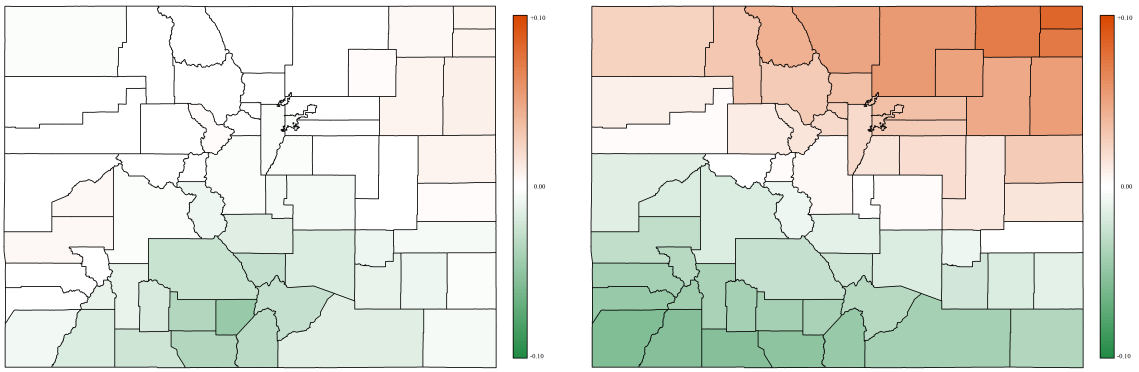


Figure 6.9: *Maps of the candidate (McCain) effects on the invalidation rate for the two methods. The GWR results are the left map; SLEM, right map. Darker oranges correspond to higher positive effects, green to higher negative effects. Shades are equivalent across maps, ranging from* $+0.10$ *to* -0.10*.*

With that said, it is interesting that the significance level is so close to α. It raises questions. Perhaps there is an interaction effect or a quadratic effect not captured in the SLEM model. Note that the northeast region also had a significant effect on the invalidation rate by the turnout. Could this effect not be sufficiently covered? Perhaps.

6.6. Conclusion

In this chapter, the second of two case-study chapters, I analyzed the 2008 Coloradoan presidential election with the expectation of no political-based unfairness. The usual analysis in Sections 6.2 through 6.4 supported this conclusion. The digit test, the regression test, and the SLEM all failed to find a statistically significant relationship between the invalidation rate and the candidate (McCain) support rate.

In Section 6.5 extended the usual geographic analysis to include additional information. As Colorado is a part of the United States, a census is taken every year ending with a zero. This census offers interesting demographic variables that can be added to the models.

In doing this, the GWR model was not a statistically significant improvement over the null model, but the SLEM model was. I then discussed the estimated effects of three

of the independent variables, turnout, poverty, and candidate support. Strictly holding to the $\alpha = 0.05$ level, the first two were significant—both statistically and practically. The effect of turnout could indicate counting fatigue. The effect of poverty could indicate issues with voter preparation.

The candidate effect was not statistically significant, but just barely. On 10,000 iterations, the estimated p-value was $\hat{p} = 0.051$. This raised an interesting question: Why is the candidate effect now (almost) significant? The temptation is to continue exploring the data, finding relationships and non-relationships. Unfortunately, once that happens, the tests become more about the data rather than the underlying data-generating-processes—a cardinal sin because the purpose of statistical analysis is to better understand the process, not just the observed data.

CHAPTER 7

CONCLUSION

In the previous chapters, I gave a firmer foundation to electoral forensics, the discipline that applies statistical techniques to election returns to determine if there is evidence of unfairness. Chapters 2 through 4 introduced, modified, and tested several statistical techniques in the hopes that they would allow us to better understand the election process.

Chapter 2 covered digit tests—tests that can be used when the government publishes vote counts at the division level. At their foundation, these were based on the original observation by Newcomb (1881) that the early pages in a book of logarithm tables wore faster than those later in the book. Benford (1938) used Newcomb's observation and showed that they could be used to distinguish between natural digit distributions and artificial ones. It is here that digit tests were born.

I examined the current Benford test and found it wanting. I generalized it by including the effect of division size and showed that tests based on the generalized Benford test were much better as statistical tests, but because the distribution of counts in a fair election is unknown, it may be dangerous to use it at this point.

So, instead of using the generalized Benford test to generate expected distributions, I tested two bootstrapping routines. The first is based on the assumption that the proportion of the vote for a given candidate (candidate support rate) follows a Logit-Normal distribution. The second is based on the non-parametric bootstrap. Both could be "tuned" so that the Type I Error rate was nominal. Neither were powerful.

Chapter 3 covered three versions of least squares regression—tests that can be used when the government gives both the vote count *and* the invalidation count. In the presence of invalidation, the free and fair hypothesis implies that the invalidation rate and the candidate support rate are independent. This chapter covered ways of allowing the computer to automatically determine the presence of two populations—fair and unfair divisions—and to determine if the resulting estimated candidate effects were significantly different from zero.

After examining three least squares methods, ordinary least squares, weighted least squares, and feasible generalized least squares, I concluded that feasible general least squares was best in terms of theory, but gave values very similar to those of weighted least squares, which did not need to iterate. Thus, I suggested weighted least squares would be an appropriate method.

In terms of detecting the existence of two populations, I tested a grid search, the Healy-Westmacott estimator (Healy and Westmacott 1956), and an empirical Bayesian method. Both the grid and the Bayesian methods gave good estimates of a threshold effect, those divisions with candidate support greater than this threshold were more likely to be unfair divisions. The Bayesian method was slower, however. The Healy-Westmacott method did not respect the threshold.

Thus, I closed Chapter 3 suggesting a combination of using weighted least squares regression and the grid search.

Chapter 4 extended its previous chapter and took geography into consideration. As it is based on people, voting is an inherently geographic process. I covered three current regression methods and proposed my own, the spatial lag expansion method (SLEM).

The current gold standard is geographically weighted regression (GWR). In comparing my new method to this, I discovered that both have poor power when dealing with maps that have too few divisions. I also discovered that the SLEM method has

poor power when dealing with non-compact maps. In all other examined cases, SLEM barely outperformed GWR.

Chapters 5 and 6 showed how to apply these techniques to test the "free and fair" claim of democratic elections. The former covered the South Sudanese Unity referendum of 2011; the latter, the 2008 Presidential election in Colorado.

The former case was straight-forward and typical of most countries in the world. The amount of information is limited to the election; censuses are rare and rarely helpful. Thus, demographic information is precious. I use Colorado to show what can be done when that information is present.

7.1. Future Work

This research did not go smoothly. However, to quote Isaac Asimov,

> The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!", but "That's funny ..."

There have been a few "That's funny" moments. Some led to answers, others to future research.

I believe I have exhausted the Benford test and its ilk (Chapter 2). Any further advancements in this area would require determining the correct "null distribution" of vote counts. I have shown that the Benford distribution is not correct. I believe I have shown that the generalized Benford distribution is also lacking. In fact, the two bootstrapping methods appear to suggest that this line of inquiry is destined for low power. With this said, I find it interesting that the generalized tests rejected the Norwegian and Irish elections. Could the geometry of the election be a factor?

In Chapter 3, I concluded that weighted least squares (WLS) was preferred to ordinary least squares (OLS) regression. This statement is misleading. Throughout that

"lost week," I struggled trying to find out why OLS performed better than WLS in my simulations. The nominal Type I Error rate for OLS was approximately 0.05, while that of WLS was approximately 0.10. This made absolutely no sense from my understanding of the underlying theory.

It turns out that William T. Dickens (1990) showed that weighted least squares, when the weights are based on populations, will rarely outperform ordinary least squares. The cause is the same issue underlying a lot of the complications in elections research: similar people cluster together. After weighting, WLS assumes that the errors are independent. "To assume these errors are independent is to assume that individuals in the same group share no common unobserved determinants" (Dickens 1990, page 329). However, this is rarely the case.

Dickens suggests the best way to determine if one should use OLS or WLS (and, by extension, FGLS) is to test the residuals for heteroskedasticity. In the way I generated the test elections, this reduced to using OLS when the division sizes were over 10,000 and WLS when they were under 1000. In real elections, one would need to use the test Dickens suggested, which I did in all cases.

Chapter 4 covered utilizing geographical information in electoral forensics. I briefly examined the effect of compactness (connectivity) and division number on the quality of the two geographical methods. That work ($n = 2$) is far from finished and promises an interesting vein of information about the relationship between the geometries and the tests.

Furthermore, the jaggedness of the power curves is very interesting to me. What causes such irregularity in the rejection rate? If we knew that, would we be able to modify the tests appropriately?

Lastly, Gaussian random fields may offer an interesting paradigm through which to view this geographical aspect.

Throughout it all, I steadfastly refused to transform the proportions using the logit transformation. Future work should be done using the transformed data, first determining which transformation is most helpful; there are several.

Finally, the most important use of this research is to *use* this research: test elections for evidence of unfairness. However, one does need to keep in mind the maxim opening this dissertation: the choice is between humility and humiliation. No matter how small the p-value or how large the effect, Type I Errors abound.

7.2. Denouement

I started this journey in the summer of 2009. I sat in the office of my then-advisor, Daniel Q. Naiman, discussing generalized linear models when I turned the conversation toward the recent election in Iran. The protesters were certain of fraud. The Western journalists were certain of fraud. It seemed as though everyone was certain of fraud. Even the Political Scientist-Statistician Walter Mebane, Jr., was certain of fraud (Mebane 2010). I could not see the evidence.

In that biweekly meeting, I discussed Mebane's evidence for it—the Benford test. Neither of us was convinced, especially given the test's origin. Professor Naiman suggested I investigate more closely. And thus was born my relationship with electoral forensics.

Almost five years later, I am at the conclusion of my first major work in this nascent field. With the statistical techniques examined in this monograph, I am better prepared to tackle the Iranian 2009 Presidential election and other elections claiming the mantle of democracy—the game is afoot.

BIBLIOGRAPHY

Afghanistan (2009a), "Polling Regulations," Internet, http://iec.org.af/pdf/legalframework/regulations/eng/RegulationOnPolling.pdf.

— (2009b), "Presidential and Provincial Council Elections," Internet, http://www.iec.org.af/results_2009/.

AFP (2010), "Constitutional body names Gbagbo I.Coast election winner," Internet, http://www.google.com/hostednews/afp/article/ALeqM5h1lqqW8eecnVcdL82ggEDWQRli0Q?docId=CNG.a5fc0e83efff72426ce88ff122d81b07.751.

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Albert, J. (2009), *Bayesian Computation with R*, Springer, 2nd ed.

Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers.

— (1995), "The Local Indicators of Spatial Association – LISA," *Geographical Analysis*, 27, 93–115.

Anselin, L., Syabri, I., and Kho, Y. (2005), "GeoDa : An Introduction to Spatial Data Analysis," *Geographical Analysis*, 38, 5–22.

BBC News (2010), "Sri Lanka Presidential Votes Being Counted," Internet, http://news.bbc.co.uk/2/hi/south_asia/8478386.stm.

— (2011), "Ivory Coast: Gbagbo held after assault on residence," Internet, http://www.bbc.co.uk/news/world-africa-13039825.

— (2012), "Sri Lanka's Sarath Fonseka Freed from Prison," Internet, http://www.bbc.co.uk/news/world-asia-18143907.

Benford, F. (1938), "The Law of Anomalous Numbers," *Proceedings of the American Philosophical Society*, 78, 551–572.

Berger, R. L. and Sinclair, D. F. (1984), "Testing Hypotheses Concerning Unions of Linear Subspaces," *Journal of the American Statistical Association*, 79, 158–163.

Bivand, R. and Yu, D. (2013), *spgwr: Geographically Weighted Regression*, R package version 0.6-22.

Carslaw, C. A. P. N. (1988), "Anomalies in Income Numbers: Evidence of Goal Oriented Behavior," *The Accounting Review*, 63, 321–327.

Carter, J. E. (2011), "Trip Report by Jimmy Carter to Sudan," Internet, http://www.cartercenter.org/news/trip_reports/sudan-011811.html.

Casella, G. and Berger, R. L. (2002), *Statistical Inference*, Wadsworth Group, 2nd ed.

Casetti, E. (1972), "Generating Models by the Expansion Method: Applications to Geographical Research," *Geographical Analysis*, 4, 81–91.

Chaturvedi, A. (1995), "A Note on the Stein Rule Estimation in Linear Models with Nonscalar Error Covariance Matrix," *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 57, 158–165.

Chen, D.-R. and Truong, K. (2012), "Using Multilevel Modeling and Geographically Weighted Regression to Identify Spatial Variations in the Relationship between Place-Level Disadvantages and Obesity in Taiwan," *Applied Geography*, 32, 737–745.

Cho, W. T. and Gaines, B. (2007), "Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance," *The American Statistician*, 61, 218–223.

Christensen, R. (2002), *Plane Answers to Complex Questions: The Theory of Linear Models*, Springer, 3rd ed.

Cliff, A. D. and Ord, J. K. (1981), *Spatial Processes: Models and Applications*, Pion Limited.

CMEV (2010), "Final Report on Election Related Violence and Malpractices: Presidential Election 2010," Internet, `http://cmev.files.wordpress.com/2010/07/presidential-election-2010-final-report.pdf`.

Collins, R. O. (2008), *A History of Modern Sudan*, Cambridge University Press.

Conover, W. J. (1999), *Practical Nonparametric Statistics*, John Wiley & Sons, 3rd ed.

Cressie, N. and Read, T. R. C. (1984), "Multinomial Goodness-of-Fit Tests," *Journal of the Royal Statistical Society. Series B (Methodological)*, 46, 440–464.

Deckert, J., Myagkov, M., and Ordeshook, P. C. (2011), "Benford's Law and the Detection of Election Fraud," *Political Analysis*, 19, 245–268.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Dickens, W. T. (1990), "Error Components in Grouped Data: Is It Ever Worth Weighting?" *The Review of Economics and Statistics*, 72, 328–333.

Epanechnikov, V. A. (1969), "Non-Parametric Estimation of a Multivariate Probability Density," *Theory of Probability and its Applications*, 14, 153–158.

Farley, R. (1839), *Tables of Logarithms*, Taylor and Walton.

Forsberg, O. J. (2012), *Terrorism and Nationalism: Theory, causes and causers*, AV Akademikerverlag, 2nd ed.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. E. (2003), *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, John Wiley & Sons.

Fotheringham, A. S., Charlton, M. E., and Brunsdon, C. (1996), "The Geography of Parameter Space: An Investigation of Spatial Non-Stationarity," *International Journal of Geographical Information Systems*, 10, 605–627.

— (1997), "Two Techniques for Exploring Non-Stationarity in Geographical Data," *Geographical Systems*, 4, 59–82.

— (1998), "Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis," *Environment and Planning A*, 30, 1905–1927.

Frederic, P. and Lad, F. (2008), "Two Moments of the Logitnormal Distribution." *Communications in Statistics: Simulation & Computation*, 37, 1263–1269.

Fromby, T. B., Johnson, S. R., and Hill, R. C. (1984), *Advanced Econometric Models*, Springer.

GADM (2014a), "GADM Shapefile: Belgium," Internet, `http://biogeo.ucdavis.edu/data/gadm2/shp/BEL\_adm.zip`.

— (2014b), "GADM Shapefile: South Sudan," Internet, `http://biogeo.ucdavis.edu/data/gadm2/shp/SSD\_adm.zip`.

— (2014c), "GADM Shapefile: Sri Lanka," Internet, `http://biogeo.ucdavis.edu/data/gadm2/shp/LKA\_adm.zip`.

Galbraith, P. W. (2009), "How the Afghan Election was Rigged," *Time Magazine*, 174.

Geary, R. C. (1954), "The Contiguity Ratio and Statistical Mapping," *The Incorporated Statistician*, 5, 115–146.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis*, Chapman & Hall, 2nd ed.

Geman, S. and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6, 721–741.

GeoCommons (2014), "GeoCommons: Colorado," Internet, `http://geocommons.com/overlays/9841.zip`.

Getis, A. and Ord, J. K. (1997), *Spatial Analysis: Modelling in a GIS Environment*, Cambridge, chap. Local Spatial Statistics: An Overview, pp. 261–278.

Ghana (2012), *C.I.75 Public Elections Regulations*, Electoral Commission of Ghana, Internet, `http://www.judicial.gov.gh/images/stories/File/C.I.%2075.pdf`.

Gullickson, A. (2007), "Linear Probability Models and Generalized Least Squares," Internet, `http://pages.uoregon.edu/aarong/teaching/G4075_Outline/node11.html`.

Guterres, J. C. (2008), "Timor-Leste: A Year of Democratic Elections," *Southeast Asian Affairs*, pp. 359–372.

Hatfield, G. D. (2011), "An Analysis of Measures of Spatial Autocorrelation," Ph.D. thesis, Oklahoma State University.

Healy, M. and Westmacott, M. (1956), "Missing Values in Experiments Analysed on Automatic Computers," *Applied Statistics*, 5, 203–206.

Hill, T. P. (1995), "A Statistical Derivation of the Significant-Digit Law," *Statistical Science*, 10, 354–363.

Holm, S. (1979), "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70.

Jeffreys, H. (1946), "An Invariant Form for the Prior Probability in Estimation Problems," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186, pp. 453–461.

Johnson, D. H. (2011), *The Root Causes of Sudan's Civil Wars: Peace or Truce*, James Currey, revised ed.

Johnson, N. L. (1949), "Systems of Frequency Curves Generated by Methods of Translation," *Biometrika*, 36, 149–176.

Keele, L. and Kelly, N. J. (2006), "Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables," *Political Analysis*, 14, 186–205.

Kennedy, P. (2003), *A Guide to Econometrics*, MIT Press, 5th ed.

Kirkpatrick, J. J. (1984), "Democratic Elections and Democratic Government," *World Affairs*, 147, pp. 61–69.

Klaassen, F. J. G. M. and Magnus, J. R. (2001), "Are Points in Tennis Independent and Identically Distributed? Evidence from a Dynamic Binary Panel Data Model," *Journal of the American Statistical Association*, 96, 500–509.

LeSage, J. P. and Pace, R. K. (2009), *Introduction to Spatial Econometrics*, Chapman & Hall.

— (2012), "The Biggest Myth in Spatial Econometrics," Internet, `http://www.wu.ac.at/wgi/en/file_inventory/lesage20120110`.

Ley, E. (1996), "On the Particular Distribution of the U.S. Stock Indexes' Digits," *The American Statistician*, 50, 311–313.

Li, S., Zhao, Z., Miaomiao, X., and Wang, Y. (2010), "Investigating spatial non-stationary and scale-dependent relationships between urban surface temperature and environmental factors using geographically weighted regression," *Environmental Modeling & Software*, 25, 1789–1800.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009), "The BUGS project: Evolution, critique, and future directions," *Statistics in Medicine*, 28, 3049–3067.

Magee, L. (1998), "Improving Survey-Weighted Least Squares Regression," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60, 115–126.

McLachlan, G. J. and Krishnan, T. (2008), *The EM Algorithm and Extensions*, Wiley-Interscience.

Mebane, Jr., W. R. (2010), "Fraud in the 2009 Presidential Election in Iran?" *Chance*, 23, 6–15.

Mebane, Jr., W. R. and Sekhon, J. S. (2004), "Robust Estimation and Outlier Detection for Overdispersed Multinomial Models of Count Data," *American Journal of Political Science*, 48, 392–411.

Murdoch, D. J., Tsai, Y.-L., and Adcock, J. (2008), "P-Values are Random Variables," *The American Statistician*, 62, 242–245.

Newcomb, S. (1881), "Note on the Frequency of Use of the Different Digits in Natural Numbers," *American Journal of Mathematics*, 4, 39–40.

Neyman, J. and Pearson, E. S. (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289–337.

Nigrini, M. (2011), *Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations*, Wiley-Corporate F&A.

— (2012), *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*, Wiley-Corporate F&A.

Ntzoufras, I. (2009), *Bayesian Modeling Using WinBUGS*, Wiley.

Paéz, A. (2005), "Local Analysis of Spatial Relationships: A Comparison of GWR and the Expansion Method," pp. 162–172.

Page, E. S. (1955), "A Test for a Change in a Parameter Occurring at an Unknown Point," *Biometrika*, 42, 523–527.

Pearson, K. R. (1900), "On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling," *Philosophical Magazine Series 5*, 50, 157–175.

Ramachandran, K. V. (1956), "On the Simultaneous Analysis of Variance Test," *The Annals of Mathematical Statistics*, 27, 521–528.

Read, T. R. C. and Cressie, N. A. C. (1988), *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer-Verlag.

Rice, J. A. (2007), *Mathematical Statistics and Data Analysis*, Brooks & Cole, 3rd ed.

Rice, S. E. (2011), "Remarks by Ambassador Susan E. Rice, U.S. Permanent Representative to the United Nations, at a Security Council Briefing on Sudan, February 9, 2011," Internet, `http://usun.state.gov/briefing/statements/2011/156244.htm`.

Sri Lanka (2010), "Presidential Election 2010," Internet, `http://www.slelections.gov.lk/presidential2010/province.html`.

Sturtz, S., Ligges, U., and Gelman, A. (2005), "R2WinBUGS: A Package for Running WinBUGS from R," *Journal of Statistical Software*, 12, 1–16.

Sudan, S. (2011), "South Sudan Statistical Yearbook 2011," Internet, `http://ssnbs.org/publications/south-sudan-statistical-yearbook-2011.html`.

US State Department (2010), "U.S. Government Statement on the Presidential Election in Sri Lanka." Internet, `http://srilanka.usembassy.gov/pr-18jan10.html`.

Wantchekon, L. (1999), "On the Nature of First Democratic Elections," *The Journal of Conflict Resolution*, 43, pp. 245–258.

Wei, W. W. S. (2006), *Time Series Analysis: Univariate and Multivariate Methods*, Addison-Wesley, 2nd ed.

Westfall, P. H. and Wolfinger, R. D. (1997), "Multiple Tests with Discrete Distributions," *The American Statistician*, 51, 3–8.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.

Wutzler, T. (2012), *`logitnorm`: Functions for the logitnormal distribution*, version 0.8.29/r32.

Yàn, M. (2010), "Gbagbo, Ouattara to enter 2nd round of presidential election in Cote d'Ivoire," Internet, `http://english.people.com.cn/90001/90777/90855/7188795.html`.

Yanagimoto, T. and Yamamoto, E. (1979), "Estimation of Safe Doses: Critical Review of the Hockey Stick Regression Method," *Environmental Health Perspectives*, 32, 193–199.

Yates, F. (1934), "Contingency Tables Involving Small Numbers and the χ^2 Test," *Supplement to the Journal of the Royal Statistical Society*, 1, 217–235.

Younger, M. S. (1979), *A Handbook for Linear Regression*, Duxbury.

VITA

Ole John Forsberg

Candidate for the Degree of

Doctor of Philosophy

Dissertation: ELECTORAL FORENSICS: TESTING THE "FREE AND FAIR" CLAIM

Major Field: Statistics

Biographical:

Education: Completed the requirements for Doctor of Philosophy in Statistics at Oklahoma State University, Stillwater, Oklahoma, in May 2014. Received a Master of Science in Engineering in Applied Mathematics and Statistics at the Johns Hopkins University, Baltimore, Maryland, in May 2010; received a Doctor of Philosophy from the University of Tennessee in Political Science; received a Master of Arts in Teaching in Secondary Education from the Johns Hopkins University, Baltimore, Maryland, in May 1992; received a Bachelor of Science in Physics and Mathematics from the University of Portland (Oregon) in 1990.

Experience: Employed as a teacher at Our Lady of Sorrows grade school in Portland, Oregon, from 1992 until 1993. Employed as a teacher at Saint Mary's Academy of Portland, Oregon, from 1994 until 2002. Employed as a graduate teaching associate at the University of Tennessee from 2002 until 2006. Employed as a lecturer at Tennessee Technological University in Cookeville, Tennessee, in 2006. Employed as a resident assistant professor at Creighton University in Omaha, Nebraska, from 2007 until 2008. Employed as a lecturer at Loyola University in Maryland from 2008 until 2010. Employed as an adjunct faculty at the University of Maryland University College from 2008 until the present.

Professional Memberships: The American Political Science Association and the American Statistical Association.